Predicting Protease Types by Hybridizing Gene Ontology and Pseudo Amino Acid Composition

Guo-Ping Zhou[1]* **and Yu-Dong Cai**[2,3]

[1]*Center for Vascular Biology Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts*
[2]*Department of Chemistry, College of Sciences, Shanghai University, Shanghai, China*
[3]*Biomedical Science Department, University of Manchester of Science and Technology, Manchester, United Kingdom*

ABSTRACT **Proteases play a vitally important role in regulating most physiological processes. Different types of proteases perform different functions with different biological processes. Therefore, it is highly desired to develop a fast and reliable means to identify the types of proteases according to their sequences, or even just identify whether they are proteases or nonproteases. The avalanche of protein sequences generated in the postgenomic era has made such a challenge become even more critical and urgent. By hybridizing the gene ontology approach and pseudo amino acid composition approach, a powerful predictor called GO-PseAA predictor was introduced to address the problems. To avoid redundancy and bias, demonstrations were performed on a dataset where none of proteins has ≥ 25% sequence identity to any other. The overall success rates thus obtained by the jackknife cross-validation test in identifying protease and nonprotease was 91.82%, and that in identifying the protease type was 85.49% among the following five types: (1) aspartic, (2) cysteine, (3) metallo, (4) serine, and (5) threonine. The high jackknife success rates yielded for such a stringent dataset indicate the GO-PseAA predictor is very powerful and might become a useful tool in bioinformatics and proteomics. Proteins 2006;63:681–684.** **© 2006 Wiley-Liss, Inc.**

Key words: gene ontology; pseudo amino acid composition; hybrid space; NN predictor; InterPro database; proteases

INTRODUCTION

Proteases play pivotal regulatory roles in the entire life cycle, namely, conception, birth, growth, digestion, maturation, aging, and death of all organisms. Proteases regulate most physiological processes by controlling the activation, synthesis, and turnover of proteins. Proteases are also essential in viruses, bacteria, and parasites for their replication and the spread of infectious diseases, in all insects, organisms, and animals for effective transmission of disease, and in human and animal hosts for the mediation and sustenance of diseases.

Knowledge of protease substrate recognition and specificity can promote identification of biologically relevant substrates, helping elucidate a protease's biological function. Also, protease protection studies are necessary to define the precise topology of the transmembrane domains of proteins in the multienzyme polysialyltransferase complex in neuroinvasive *E. coli* K1.[1–4]

The importance of proteases by nature has made them become a focused target of drug design (see, e.g., Refs. [5–22] as well as a recent review[23]). As is well known, the actions of proteases are highly selective, with each protease being responsible for splitting very specific sequences of amino acids under a preferred set of environmental conditions. The rapidly increasing number of protein sequences entering into data banks has called for development of automated methods to address the two very important yet quite practical problems: (1) How can we fast identify if it is a protease or nonprotease for a newly found protein sequence? (2) Is it possible to predict which type it belongs to for an uncharacterized protease according to its sequence information?

MATERIALS AND METHOD

First, we need an unbiased training dataset, which is obtained via the following procedures: (1) The classification of protease types was based on the MEROPS database (release 6.20, 24 March 2003) at http://merops.sanger.ac.uk/ and the corresponding sequences were obtained from the databases of UniProt/Swiss-Prot at http://www.ebi.ac.uk/swissprot (Release 44, 5 July 2004) and UniProt/TrEMBL at http://www.ebi.ac.uk/trembl (Release 27.0, 5 July 2004). (2) Those sequences which are less than 50 amino acids in length were removed because they might not represent the entire sequence but just some fragments. (3) To avoid any homologous bias, a redundancy cutoff was imposed by PISCES[24] to exclude those sequences that have ≥ 25% sequence identity to any other in a same subset.

Thus, a total of 510 protease sequences were generated that consist of 42 aspartic proteases, 95 cysteine proteases, 216 metallo proteases, 141 serine proteases, and 16 threonine proteases. Meanwhile, by following the same steps, a

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Guo-Ping Zhou, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Division of Molecular and Vascular Medicine, E/RW 759, Boston, MA, 02115. E-mail: gzhou@bidmc.harvard.edu

Received 19 October 2005; Revised 8 November 2005; Accepted 8 November 2005

Published online 2 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20898

total of 652 nonprotease protein sequences were randomly taken from the UniProt/Swiss-Prot databank as well. The accession numbers of the 510 protease proteins (classified into 5 types) and the 652 nonprotease proteins are given in the Supplementary Material.

By following the procedures elaborated in Chou and Cai,[25,26] the protein samples studied here can be represented in terms of the 1930D (dimensional) GO space as given below:

$$\mathbf{P} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_j \\ \vdots \\ g_{1930} \end{bmatrix}, \qquad (1)$$

where $g_j = 1$ if there is a hit corresponding to the jth ($j = 1, 2, \ldots, 1930$) GO number[25,26] when using the program IPRSCAN[27] to search InterPro functional domain database (release 6.1)[27] for the protein **P**; otherwise, $g_j = 0$. The detailed procedure in defining the1930D GO space and the discussion of its advantage can be found in Chou and Cai.[28,29]

In case no such a hit whatsoever was found, the protein **P** formulated by Equation 1 will correspond to a naught vector. To cope with such a circumstance, the protein is instead defined in the (20 + λ)D PseAA space,[30] as given below:

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix}, \qquad (2)$$

where $p_1, p_2, \ldots, p_{20}$ represent the 20 components of the classical amino acid composition,[31–33] while $p_{20\,+\,1}$ is the first-tier sequence order correlation factor, $p_{20\,+\,2}$ the second-tier sequence order correlation factor, and so forth (see Fig. 1 of Chou,[34] or Fig. 2 of Chou[35]). It is the additional λ components in Equation 2 that incorporate some sequence-order effects into the representation of a protein sample. The detailed procedure in how to select the optimal value for λ is given in Chou.[30] In the current study, the optimal value for λ is 31. Given a protein, the (20 + 31) = 51 PseAA components in Equation 2 can be easily derived by following the procedures as described by Chou.[30,35] Thus, any protein that corresponds to a naught vector in the 1930D GO space (Eq. 1) can always be uniquely defined in the 51D PseAA space (Eq. 2).

The prediction was performed with the nearest neighbor (NN) algorithm.[36,37] The NN predictor is particularly useful for the situation when the distributions of the samples are unknown. During the course of prediction, the following self-consistency principle should be followed: If a query protein was defined in the 1930D GO space (Eq. 1), then the prediction should be conducted based on those proteins in the training set that could also be defined in the same 1930D space. If the query protein in the 1930D GO space was a naught vector and hence must be defined instead in the (20 + λ)D PseAA composition space (see Eq. 2), then the prediction should be conducted according to the principle that all the proteins in the training set be defined in the same (20 + λ)D PseAA composition space as well. Accordingly, the current NN predictor actually consists of two subpredictors: (1) the NN-1930D GO predictor that operates in the 1930D GO space, and (2) the NN-51D PseAA predictor that operates in the 51D pseudo amino acid composition space with λ = 31. The entire process is called GO-PseAA hybridization approach.

TABLE I. Breakdown of the Protein Entries into the Subset Defined in the 1930D GO Space (Eq. 1) and that in the 51D PseAA Composition Space (Eq. 2)

Dataset[a]	1930D GO Space	51D PseAA Space	Total
Protease	462	48	510
Nonprotease	594	58	652

[a]From the Supplementary Material.

TABLE II. Success Rates in Identifying Protease and Nonprotease Proteins by the Jackknife Cross-Validation Test[a]

Protease	Nonprotease	Overall
$\frac{479}{510} = 93.92\%$	$\frac{588}{652} = 90.18\%$	$\frac{1067}{1162} = 91.82\%$

[a]Using the data of the Supplementary Materials to perform the jackknife cross-validation test.

RESULTS AND DISCUSSION

For the proteins listed in the Supplementary Material, we obtained the following results according to procedures described in the Materials and Methods section: (1) Of the 510 protease sequences, 462 got the hits and hence were defined in the 1930D GO space, and the remainder defined in the 51D PseAA space (Table I). (2) Of the 652 nonprotease sequences, 594 were defined in the 1930D GO space, and the remainder defined in the 51D PseAA space. This means that, if the definition of proteases was only based on GO approach, 510 − 462 = 48 proteins in the protease set and 652 − 594 = 58 proteins in the nonprotease set would have no definition, leading to a failure of identifying their attribute. That is why it is so important to hybridize with the PseAA approach, by which not only a protein can always be defined but also its sequence-order effects may considerably be taken into account.[30]

To show the power of the current approach, the jackknife cross-validation test was performed on the dataset in the Supplementary Material. As is well known, the single independent dataset test, subsampling test, and jackknife test are the three procedures often used for cross-validation in statistical prediction. Of these three, the jackknife test is regarded as the most objective and effective one as elucidated in a comprehensive review.[38]

TABLE III. Success Rates in Identifying Protease Types by the Jackknife Cross-Validation Test

Aspartic	Cysteine	Metallo	Serine	Threonine	Overall
$\frac{34}{42} = 80.95\%$	$\frac{81}{95} = 85.26\%$	$\frac{194}{216} = 89.81\%$	$\frac{114}{141} = 80.85\%$	$\frac{13}{16} = 81.25\%$	$\frac{436}{510} = 85.49\%$

[a]See footnote a of Table II for further explanation.

Recently, the jackknife test has been used by more and more investigators to examine the power of various prediction methods (see, e.g., Refs. 28, 31, 34, 39–53). Accordingly, the real power of a predictor should be measured by the success rate of jackknife test. As shown in Tables II and III, the overall jackknife success rates obtained by the current GO-PseAA hybridization approach are 91.82% for the case between protease and nonprotease, and 85.49% for the case among the five protease types. These rates are very high for such a stringent dataset in which none of proteins has ≥ 25% sequence identity to any others.

CONCLUSION

Hybridizing the gene ontology approach (GO) with the pseudo amino acid composition approach (PseAA) can make the two powerful approaches complement each other in grasping the sequence pattern feature for identifying protease types. Particularly, it can make allowance for bringing out the best in each other and making each shining more brilliant in the other's company.

With the avalanche of protein sequences we are facing in the postgenomic era, the current computational method may become a useful high throughput tool in bridging the huge gap between the number of sequence-known proteins and the number of function-known proteins.

ACKNOWLEDGMENTS

The authors wish to thank the three anonymous reviewers whose constructive suggestions were very helpful for strengthening the presentation of this article.

REFERENCES

1. Pigeon RP, Silver RP. Topological and mutational analysis of KpsM, the hydrophobic component of the ABC-transporter involved in the export of polysialic acid in *Escherichia coli* K1. Mol Microbiol 1994;14:871–881.
2. Zhou GP, Troy FA. NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. Curr Protein Peptide Sci 2005;6:399–411.
3. Zhou GP, Troy FA 2nd. Characterization by NMR and molecular modeling of the binding of polyisoprenols and polyisoprenyl recognition sequence peptides: 3D structure of the complexes reveals sites of specific interactions. Glycobiology 2003;13:51–71.
4. Zhou GP, Troy FA 2nd. NMR study of the preferred membrane orientation of polyisoprenols (dolichol) and the impact of their complex with polyisoprenyl recognition sequence peptides on membrane structure. Glycobiology 2005;15:347–359.
5. Poorman RA, Tomasselli AG, Heinrikson RL, Kezdy FJ. A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J Biol Chem 1991;266:14554–14561.
6. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 1993;268:16938–16948.
7. Chou KC. Review: prediction of HIV protease cleavage sites in proteins. Anal Biochem 1996;233:1–14.
8. Chou JJ. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J Protein Chem 1993;12:291–302.
9. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 1993;268:6119–6124.
10. Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 1993;268:14875–14880.
11. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 1993;32:6548–6554.
12. Chou KC, Kezdy FJ, Reusser F. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 1994;221:217–230.
13. Du QS, Wang SQ, Wei DQ, Zhu Y, Guo H, Sirois S, Chou KC. Polyprotein cleavage mechanism of SARS CoV Mpro and chemical modification of octapeptide. Peptides 2004;25:1857–1864.
14. Sirois S, Wei DQ, Du QS, Chou KC. Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. J Chem Inf Comput Sci 2004;44:1111–1122.
15. Du QS, Wang S, Wei DQ, Sirois S, Chou KC. Molecular modelling and chemical modification for finding peptide inhibitor against SARS CoV Mpro. Anal Biochem 2005;337:262–270.
16. Du QS, Wang SQ, Jiang ZQ, Gao WN, Li YD, Wei DQ, Chou KC. Application of bioinformatics in search for cleavable peptides of SARS-CoV Mpro and chemical modification of octapeptides. Medicinal Chem 2005;1:209–213.
17. Gan YR, Huang H, Huang YD, Rao CM, Zhao Y, Liu JS, Wu L, Wei DQ. Synthesis and activity assess of an octapeptide inhibitor designed for sars coronavirus main proteinase. Peptides 2005. Forthcoming.
18. Chou KC, Jones D, Heinrikson RL. Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett 1997;419:49–54.
19. Chou JJ, Matsuo H, Duan H, Wagner G. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. Cell 1998;94:171–180.
20. Chou KC, Tomasselli AG, Heinrikson RL. Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett 2000;470:249–256.
21. Chou KC, Howe WJ. Prediction of the tertiary structure of the beta-secretase zymogen. Biochem Biophys Res Commun 2002;292:702–708.
22. Chou KC, Wei DQ, Zhong WZ. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS [Erratum: 2003;310:675]. Biochem Biophys Res Comm 2003;308:148–151.
23. Chou KC. Review: structural bioinformatics and its impact to biomedical science. Curr Medicinal Chem 2004;11:2105–2134.
24. Wang GL, Dunbrack Jr. RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
25. Chou KC, Cai YD. Predicting enzyme family class in a hybridization space. Protein Sci 2004;13:2857–2863.
26. Chou KC, Cai YD. Using GO-PseAA predictor to identify membrane proteins and their types. Biochem Biophys Res Commun 2005;327:845–847.

27. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR and others. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 2001;29:37–40.
28. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765–45769.
29. Chou KC, Cai YD. Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 2004; 321:1007–1009.
30. Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 2001;43:246–255.
31. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
32. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–344.
33. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–22020.
34. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21:10–19.
35. Chou KC. Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Peptide Sci 2005;6:423–436.
36. Friedman JH, Baskett F, Shustek LJ. An algorithm for finding nearest neighbors. IEEE Trans Inform Theor 1975;C-24:1000–1006.
37. Cai YD, Chou KC. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Commun 2003;305:407–411.
38. Chou KC, Zhang CT. Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
39. Chou KC, Cai YD. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 2004;91:1197–1203.
40. Chou KC. Prediction of G-protein-coupled receptor classes. J Proteome Res 2005;4:1413–1418.
41. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins 2001;44:57–59.
42. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins 2003;50:44–48.
43. Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 2003;22:395–402.
44. Luo RY, Feng ZP, Liu JK. Prediction of protein strctural class by amino acid and polypeptide composition. Eur J Biochem 2002;269: 4219–4225.
45. Feng ZP. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 2001;58:491–499.
46. Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC. Using complexity measure factor to predict protein subcellular location. Amino Acids 2005;28:57–61.
47. Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids DOI 10.1007/s00726-005-0225-6 2005.
48. Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 2005;28:373–376.
49. Wang M, Yang J, Liu GP, Xu ZJ, Chou KC. Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Design Selection 2004;17:509–516.
50. Wang M, Yang J, Xu ZJ, Chou KC. SLLE for predicting membrane protein types. J Theor Biol 2005;232:7–15.
51. Feng KY, Cai YD, Chou KC. Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 2005;334:213–217.
52. Shen HP, Yang J, Liu XJ, Chou KC. Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 2005;334:577–581.
53. Shen H, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 2005;334: 288–292.