Automatic Labeling of Protein Structures by Hidden Markov Models

Kiyoshi Asai[1] asai@etl.go.jp
Kentaro Onizuka[2] onizuka@mrit.mei.co.jp
Masayuki Akahoshi[3] akahoshi@icot.or.jp
Hidetoshi Tanaka[3] htanaka@icot.or.jp
Katunobu Itou[1] kito@etl.go.jp

[1] Electrotechnical Laboratory (ETL), 1-1-4 Umezono, Tsukuba, 305 Japan
[2] Matsushita Research Institute, 3-10-1 Higashimita, Tama-ku, Kawasaki, 214 Japan
[3] ICOT, Mita Kokusai Bldg. 21F, 1-4-28 Mita, Minato-ku, Tokyo, 108 Japan

In this research, local structure labeling of protein is performed by Hidden Markov Models (HMMs) using Multi Scale Structure Description (MSSD).

HMMs have been used for structure prediction [Asai91, Asai93A, Asai93B], for sequence alignment [Haussler93], for protein classification [Tanaka93], and for motif extraction [Fujiwara94]. Most of them used 20 amino acids as the discrete output symbols of distributions in HMMs. In this paper, however, HMMs have continuous output distributions for the hidden states, which output MSSD-parameters of the protein structures. MSSD is a robust parameterization of protein structures using 3D coordinates of α carbons [Onizuka94].

In order to get appropriate HMMs for the purpose, the network shapes of the HMMs must be determined. The HMM training here consists of parameter learning and of dynamic network shape growth. For the network shape determination, iterative duplication method [Fujiwara94] and successive state splitting (SSS) algorithm [Tanaka93] have been used for protein HMMs. We used modified iterative duplication method, where negligible links are deleted and states of the largest output variances are duplicated.

After the HMM training, not only the output distributions of the states, but also the transition probabilities between the states characterize the features of the local structures. Therefore, both continuous structures and short ranged structures are categorized naturally as the hidden states in HMMs. By using the parameters of 5-residue MSSD, which nearly correspond to the secondary structures, "α helix" and "β strand" and many types of "turns" and "coils" are expressed in HMMs.

[1]浅井 潔、伊藤克亘：電子技術総合研究所知能情報部音声研究室, 〒 305 つくば市梅園 1-1-4
[2]鬼塚健太郎：(株) 松下技研ヒューマンインタフェース研究所, 〒 214 川崎市多摩区東三田 3-10-1
[3]赤星正幸、田中秀俊：新世代コンピュータ技術開発機構, 〒 108 港区三田 1-4-28 国際ビル 21F

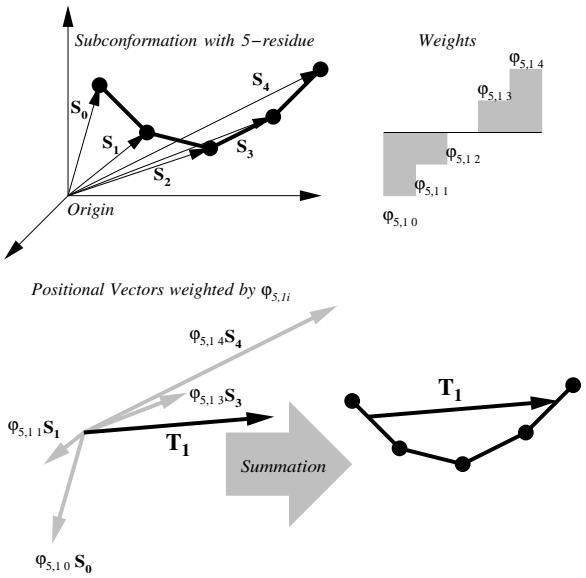

Fig.1: Abstraction of Structures in MSSD

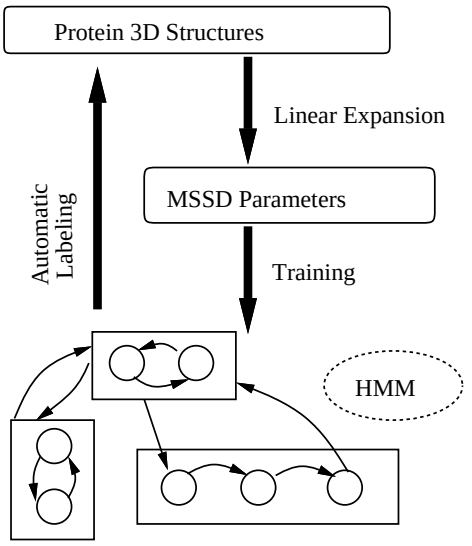

Fig.2: Labeling of Structures by HMMs

By estimation of the hidden state transitions using Viterbi algorithm, the protein structures are aligned to the HMMs. The labeling of the local structures is an easy translation of this alignment. At the same time, the HMMs extract the rules between the local structures as the matrices of the transition probabilities. These rules are important for modeling the protein structures including the higher level structures.

References

[Asai91] Asai,K.; Hayamizu,S. and Handa,K.: "Genetic Information Processing by Stochastic Model: HMM for Secondary Structure Prediction of Protein," *Proc. Genome Info. Workshop 2, pp.144-147. (in Japanese, 1991).*

[Asai93A] Asai,K.; Hayamizu,S. and Handa,K.: "Secondary Structure Prediction by Hidden Markov Model," *CABIOS Vol.9, No.2, pp.141-146, 1993.*

[Asai93B] Asai,K.; Hayamizu,S. and Onizuka,K.: "HMM with Protein Structure Grammar," *Proc. HICSS26, vol.1, pp.783-791.(1993).*

[Fujiwara94] Fujiwara,Y.; Asogawa,M. and Konagaya,A.: "Stochastic Motif Extraction Using Hidden Markov Model," *Proc. ISMB94, pp.121-129.(1994).*

[Haussler93] Haussler,D.;Krogh,A.;Mian,I.S. and Sjölander,K.: "Protein modeling using hidden Markov Models: Analysis of globins," *Proc. HICSS26, vol.1, pp.792-802. (1993).*

[Onizuka94] Onizuka,K.; Akahoshi,M.; Ishikawa,M. and Asai,K.: "The Multi-Scale 3D-1D Compatibility Scoring for Inverse Protein Folding Problem," *Proc. of ISMB-94 pp.314-321. (1994).*

[Tanaka93] Tanaka,H.; Onizuka,K. and Asai,K.: "Classification of Proteins via Successive State Splitting of Hidden Markov Network," *Proc. W26 in IJCAI93. (1993).*