1: Curr Opin Struct Biol. 1996 Jun;6(3):361-5.

Hidden Markov models.

Eddy SR.

Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA. eddy@genetics.wustl.edu

'Profiles' of protein structures and sequence alignments can detect subtle homologies. Profile analysis has been put on firmer mathematical ground by the introduction of hidden Markov model (HMM) methods. During the past year, applications of these powerful new HMM-based profiles have begun to appear in the fields of protein-structure prediction and large-scale genome-sequence analysis.

Publication Types:
    Research Support, Non-U.S. Gov't
    Research Support, U.S. Gov't, P.H.S.
    Review

PMID: 8804822 [PubMed - indexed for MEDLINE]

2: DNA Res. 1997 Jun 30;4(3):179-84.

Prediction of translation initiation sites on the genome of Synechocystis sp. strain PCC6803 by Hidden Markov model.

Hirosawa M, Sazuka T, Yada T.

Kazusa DNA Research Institute, Chiba, Japan. hirosawa@kazusa.ur.jp

We developed a computer program, GeneHackerTL, which predicts the most probable translation initiation site for a given nucleotide sequence. The program requires that information be extracted from the nucleotide sequence data surrounding the translation initiation sites according to the framework of the Hidden Markov Model. Since the translation initiation sites of 72 highly abundant proteins have already been assigned on the genome of Synechocystis sp. strain PCC6803 by amino-terminal analysis, we extracted necessary information for GeneHackerTL from the nucleotide sequence data. The prediction rate of the GeneHackerTL for these proteins was estimated to be 86.1%. We then used GeneHackerTL for prediction of the translation initiation sites of 24 other proteins, of which the initiation sites were not assigned experimentally, because of the lack of a potential initiation codon at the amino-terminal position. For 20 out of the 24 proteins, the initiation sites were predicted in the upstream of their amino-terminal positions. According to this assignment, the processed regions represent a typical feature of signal peptides. We could also predict multiple translation initiation sites for a particular gene for which at least two initiation sites were experimentally detected. This program would be effective for the prediction of translation initiation sites of other proteins, not only in this species but also in other prokaryotes as well.

Publication Types:
    Research Support, Non-U.S. Gov't

PMID: 9330905 [PubMed - indexed for MEDLINE]

3: DNA Res. 1997 Feb 28;4(1):1-7.

Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome using the hidden Markov model.

Yada T, Sazuka T, Hirosawa M.

Japan Science and Technology Corporation (JST), Tokyo, Japan.
yada@tokyo.jstc.go.jp

Sequence patterns surrounding the translation initiation sites of Cyanobacterium were precisely analyzed by the hidden Markov model (HMM) based on the actual translation initiation sites. In a previous study, 72 actual protein coding regions and their translation initiation sites on the genome of Synechocystis sp. strain PCC6803 were determined by Sazuka et al. using protein two-dimensional electrophoresis and microsequening. In this work, we extracted the sequence patterns surrounding translation initiation sites as HMM using the computer program YEBIS. The constructed HMM could recognize all but one translation initiation site. The HMM contains an AG-rich region (5.7 bp on average), as the Shine-Dalgarno sequence exclusively contains purines, upstream of the translation initiation site (-9.7 position on average) and a CT rich region (4.2 bp on average) just upstream from the translation initiation site. In addition, we found that the second amino acid (-4.5,6) could be classified into two types, one of which had C as their second codon while another of which has a nucleotide distribution relatively similar to the distribution among amino acids in the 72 proteins. This fact corresponds well to our earlier finding that when the second nucleotide of the second amino acid of a translated protein was C, an initial methionine was processed and that otherwise the methionine was intact with high frequency.

Publication Types:
    Research Support, Non-U.S. Gov't

PMID: 9179490 [PubMed - indexed for MEDLINE]

4: DNA Res. 1996 Dec 31;3(6):355-61.

Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model.

Yada T, Hirosawa M.

Japan Science and Technology Corporation (JST), Tokyo, Japan.

The gene-finding programs developed so far have not paid much attention to the detection of short protein coding regions (CDSs). However, the detection of short CDSs is important for the study of photosynthesis. We utilized GeneHacker, a gene-finding program based on the hidden Markov model (HMM), to detect short CDSs (from 90 to 300 bases) in a 1.0 mega contiguous sequence of cyanobacterium Synechocystis sp. strain PCC6803 which carries a complete set of genes for oxygenic photosynthesis. GeneHacker differs from other gene-finding programs based on the HMM in that it utilizes di-codon statistics as well. GeneHacker successfully detected seven out of the eight short CDSs annotated in this sequence and was clearly superior to GeneMark in this range of length. GeneHacker detected 94 potentially new CDSs, 9 of which have counterparts in the genetic databases. Four of the nine CDSs were less than 150 bases and were

photosynthesis-related genes. The results show the effectiveness of GeneHacker in detecting very short CDSs corresponding to genes.

Publication Types:
Research Support, Non-U.S. Gov't

PMID: 9097038 [PubMed - indexed for MEDLINE]

5: J Comput Biol. 1997 Fall;4(3):311-23.

Improved splice site detection in Genie.

Reese MG, Eeckman FH, Kulp D, Haussler D.

Human Genome Informatics Group, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. mgreese@lbl.gov

We present an improved splice site predictor for the genefinding program Genie. Genie is based on a generalized Hidden Markov Model (GHMM) that describes the grammar of a legal parse of a multi-exon gene in a DNA sequence. In Genie, probabilities are estimated for gene features by using dynamic programming to combine information from multiple content and signal sensors, including sensors that integrate matches to homologous sequences from a database. One of the hardest problems in genefinding is to determine the complete gene structure correctly. The splice site sensors are the key signal sensors that address this problem. We replaced the existing splice site sensors in Genie with two novel neural networks based on dinucleotide frequencies. Using these novel sensors, Genie shows significant improvements in the sensitivity and specificity of gene structure identification. Experimental results in tests using a standard set of annotated genes showed that Genie identified 86% of coding nucleotides correctly with a specificity of 85%, versus 80% and 84% in the older system. In further splice site experiments, we also looked at correlations between splice site scores and intron and exon lengths, as well as at the effect of distance to the nearest splice site on false positive rates.

Publication Types:
Research Support, U.S. Gov't, Non-P.H.S.

PMID: 9278062 [PubMed - indexed for MEDLINE]

6: Mol Biol Evol. 1996 Jan;13(1):93-104.

A Hidden Markov Model approach to variation among sites in rate of evolution.

Felsenstein J, Churchill GA.

Department of Genetics, University of Washington, Seattle 98195, USA.

The method of Hidden Markov Models is used to allow for unequal and unknown evolutionary rates at different sites in molecular sequences. Rates of evolution at different sites are assumed to be drawn from a set of possible rates, with a finite number of possibilities. The overall likelihood of phylogeny is calculated as a sum of terms, each term being the probability of the data given a particular assignment of rates to sites, times the prior probability of that particular combination of rates. The probabilities of different rate combinations are specified by a stationary Markov chain that assigns rate categories to sites. While there will be a very large number of possible ways of

assigning rates to sites, a simple recursive algorithm allows the contributions to the likelihood from all possible combinations of rates to be summed, in a time proportional to the number of different rates at a single site. Thus with three rates, the effort involved is no greater than three times that for a single rate. This "Hidden Markov Model" method allows for rates to differ between sites and for correlations between the rates of neighboring sites. By summing over all possibilities it does not require us to know the rates at individual sites. However, it does not allow for correlation of rates at nonadjacent sites, nor does it allow for a continuous distribution of rates over sites. It is shown how to use the Newton-Raphson method to estimate branch lengths of a phylogeny and to infer from a phylogeny what assignment of rates to sites has the largest posterior probability. An example is given using beta-hemoglobin DNA sequences in eight mammal species; the regions of high and low evolutionary rates are inferred and also the average length of patches of similar rates.

Publication Types:
    Research Support, U.S. Gov't, Non-P.H.S.
    Research Support, U.S. Gov't, P.H.S.

PMID: 8583911 [PubMed - indexed for MEDLINE]

7: J Mol Biol. 2004 May 14;338(5):1027-36.

A combined transmembrane topology and signal peptide prediction method.

Kall L, Krogh A, Sonnhammer EL.

Center for Genomics and Bioinformatics, Karolinska Institutet, SE-17 177 Stockholm, Sweden.

An inherent problem in transmembrane protein topology prediction and signal peptide prediction is the high similarity between the hydrophobic regions of a transmembrane helix and that of a signal peptide, leading to cross-reaction between the two types of predictions. To improve predictions further, it is therefore important to make a predictor that aims to discriminate between the two classes. In addition, topology information can be gained when successfully predicting a signal peptide leading a transmembrane protein since it dictates that the N terminus of the mature protein must be on the non-cytoplasmic side of the membrane. Here, we present Phobius, a combined transmembrane protein topology and signal peptide predictor. The predictor is based on a hidden Markov model (HMM) that models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of interconnected states. Training was done on a newly assembled and curated dataset. Compared to TMHMM and SignalP, errors coming from cross-prediction between transmembrane segments and signal peptides were reduced substantially by Phobius. False classifications of signal peptides were reduced from 26.1% to 3.9% and false classifications of transmembrane helices were reduced from 19.0% to 7.7%. Phobius was applied to the proteomes of Homo sapiens and Escherichia coli. Here we also noted a drastic reduction of false classifications compared to TMHMM/SignalP, suggesting that Phobius is well suited for whole-genome annotation of signal peptides and transmembrane regions. The method is available at as well as at

Publication Types:
    Research Support, Non-U.S. Gov't

PMID: 15111065 [PubMed - indexed for MEDLINE]

8: J Mol Biol. 2001 Jan 19;305(3):567-80.

Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL.

Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, 2800 Lyngby, Denmark. krogh@cbs.dtu.dk

We describe and validate a new membrane protein topology prediction method, TMHMM, based on a hidden Markov model. We present a detailed analysis of TMHMM's performance, and show that it correctly predicts 97-98 % of the transmembrane helices. Additionally, TMHMM can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99 %, although the accuracy drops when signal peptides are present. This high degree of accuracy allowed us to predict reliably integral membrane proteins in a large collection of genomes. Based on these predictions, we estimate that 20-30 % of all genes in most genomes encode membrane proteins, which is in agreement with previous estimates. We further discovered that proteins with N(in)-C(in) topologies are strongly preferred in all examined organisms, except Caenorhabditis elegans, where the large number of 7TM receptors increases the counts for N(out)-C(in) topologies. We discuss the possible relevance of this finding for our understanding of membrane protein assembly mechanisms. A TMHMM prediction service is available at http://www.cbs.dtu.dk/services/TMHMM/. Copyright 2001 Academic Press.

Publication Types:
    Research Support, Non-U.S. Gov't

PMID: 11152613 [PubMed - indexed for MEDLINE]

9: Proc Int Conf Intell Syst Mol Biol. 1998;6:175-82.

A hidden Markov model for predicting transmembrane helices in protein sequences.

Sonnhammer EL, von Heijne G, Krogh A.

National Center for Biotechnology Information, NLM/NIH, Bethesda, Maryland 20894, USA. esr@ncbi.nlm.nih.gov

A novel method to model and predict the location and orientation of alpha helices in membrane-spanning proteins is presented. It is based on a hidden Markov model (HMM) with an architecture that corresponds closely to the biological system. The model is cyclic with 7 types of states for helix core, helix caps on either side, loop on the cytoplasmic side, two loops for the non-cytoplasmic side, and a globular domain state in the middle of each loop. The two loop paths on the non-cytoplasmic side are used to model short and long loops separately, which corresponds biologically to the two known different membrane insertions mechanisms. The close mapping between the biological and computational states allows us to infer which parts of the model architecture are important to capture the information that encodes the membrane topology, and to gain a better understanding of the mechanisms and constraints involved. Models were estimated both by maximum likelihood and a discriminative method, and a method for reassignment of the membrane helix boundaries were developed.

In a cross validated test on single sequences, our transmembrane HMM, TMHMM, correctly predicts the entire topology for 77% of the sequences in a standard dataset of 83 proteins with known topology. The same accuracy was achieved on a larger dataset of 160 proteins. These results compare favourably with existing methods.

Publication Types:
    Research Support, Non-U.S. Gov't

PMID: 9783223 [PubMed - indexed for MEDLINE]