서열분석알고리즘 HMM관련 프로젝트

각자의 취향에 따라 다음 두개의 프로젝트중 하나를 선택하여 수행한 후 담담조교의 이 메일로 제출해주십시오. 문의사항은 각 문제 담당조교에게 문의해 주십시오.

Project 1) HMM을 구현하여 작은 규모의 Sequence 분석.
(담당조교: 엄재홍, jheom@bi.snu.ac.kr)
Project 2) 기존의 HMM 기반 tool을 사용한 대규모의 Sequence 분석.
(담당조교: 지성욱, swchi@bi.snu.ac.kr)

Project 1)

생물학 관련 문제들중 이미 답이 알려져 있는 문제 (예를 들어 Exon,Intron 구별문제, Promoter인식등) 중에서 자신이 관심이 있는 문제를 골라서 HMM model을 다음 조건에 맞추어서 구현하여 결과를 제출해주십시오.

1) 다음 논문을 Tutorial로 생각하시고 HMM model을 작성해주십시오
Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77:257-286.
(다음 논문은 프로젝트 게시판에서 다운받으실수 있습니다.)

2) 사용코드 종류나 실행환경에 대한 제한은 없으며 제출시 Source code에 주석을 달아 제출해 주시고 각각의 실행환경과 실행법을 설명하여 제출해 주십시오.

3) Training Set와 Test Set의 크기는 방대할 필요는 없으나 너무 적으면 감점이 되므로 적당한 사이즈로 선택해주시고 각각의 set에 대한 설명을 제출해 주십시오

4) HMM model 작성후의 실험결과는 training의 확률결과나 인식결과등으로 나타내서 제출해주시기 바랍니다.

평가기준
1) Code 작성
2) 프로그램에 대한 data tuning 정도와 performace.
3) 프로그램 주석에 대한 평가 (HMM model에 대한 이해정도)

Project 2)

HMM 기반 tool중 HMMER를 사용하여 우선은 Tutorial을 실행해서 tool을 이해한 후에 다음 7개의 연습문제를 풀어 제출해주시기 바랍니다.

1) 제출시에는 각각의 연습문제시 생성된 파일과 각 연습문제에 대한 설명 및 실행결과를 제출해 주십시오.
2) 연습문제에서 제시된 C2 domain외에 다른 도메인을 가지고 프로젝트를 수행하실때에는 가산점이 주어지며 사용한 도메인에 대한 설명과 각 연습문제 단계별로 간단한 설명을 첨부해서 제출해 주십시오.
3) 평가기준: HMM model 기반 tool에 대한 이해와 이 tool을 응용한 생물학 문제에 대한 해결능력.

Tutorial

1. Install HMMER2.2

Install HMMER 2.2 in your path, HMMER's on-line documentation and a tutorial introduction to the functions in the package is here (http://hmmer.wustl.edu)

2. Practice Tutorial.

The Tutorial should be sufficient to get you started on work of these exercises.
On-line Tutorial (http://hmmer.wustl.edu/hmmer-html/node2.html)
(Tutorial 은 프로젝트 게시판에서 다운받으실 수 있습니다)

Exercise

1. You have to make a multiple sequence alignment of a protein domain you're interested in or C2 domain 1.
   Hint) You can use freely available multiple sequence alignment programs such as CLUSTALW or just download from pfam database (http://pfam.wustl.edu/).

2. Use the program **hmmbuild** to build an HMM from an alignment.

3. Search the **SWISS-PROT** with your own *hmm* (using **hmmsearch).** Look at the E() values for the high scoring protein in SWISS-PROT. You can download **SWISS-PROT** from this site (http://www.expasy.ch/sprot/)

4. Use the program **hmmcalibrate** to determine some statistics for your new HMM, so that HMMER can estimate E-values fairly accurately in any subsequent searches you do with the HMM.

5. Creating your own profile HMM database. Build an HMM database called ***myhmms*** that contains models of C2 domain and the pkinase protein kinase catalytic domain (/tutorial/pkinase.slx)

6. Parsing the domain structure of a sequence with **hmmpfam.** Use myhmms to analyze the *Schizosaccharomyces pombe Protein Kinase C-like 1.(* PCK1_SCHPO (P36582)).

7. Search for domains of *Schizosaccharomyces pombe Protein Kinase C-like 1* in PFAM database. You can Download the PFAM database from here. (http://www.sanger.ac.uk/pfam/)

* 자신이 관심있는 도메인을 가지고 프로젝트를 수행할 경우에는 5번 연습문제에서는 그 도메인을 포함하여 적어도 2개이상의 다른 도메인이 포함된 HMM database를 구성하고 6번 7번에서는 protein kinase C-like 1 대신에 그 두 도메인을 포함하는 protein sequence를 가지고 분석하기 바랍니다.

C2 domain 1

CNE1_HUMAN (Q99829), CNE3_HUMAN (O75131), KPC1_APLCA (Q16974), KPC1_DROME (P05130), KPC1_HUMAN (P05771), KPC1_LYTPI (Q25378), KPC1_RABIT (P05772), KPC1_RAT (P04410), KPC2_BOVIN (P05126), KPC2_DROME (P13677), KPC2_HUMAN (P05127), KPC2_MOUSE (P04411), KPC2_RABIT (P05773), KPCA_BOVIN (P04409), KPCA_HUMAN (P17252), KPCA_MOUSE (P20444), KPCA_RABIT (P10102), KPCA_RAT (P05696), KPCG_BOVIN (P05128), KPCG_HUMAN (P05129), KPCG_MOUSE (P05697), KPCG_RABIT (P10829), NED4_HUMAN (P46934), NED4_MOUSE (P46935), PERF_HUMAN (P14222), PERF_MOUSE (P10820), PERF_RAT (P35763), PUB1_SCHPO (Q92462), RSP5_YEAST (P39940), SUF1_HUMAN (Q9HCE7), SUF1_XENLA (Q9PUN2), SUF2_HUMAN (Q9HAU4), SYT4_HUMAN (Q9H2B2), UN13_CAEEL (P27715), YGJI_CAEEL (Q9XUB9), YMH2_YEAST (Q03640), RP3A_BOVIN (Q06846), RP3A_HUMAN (Q9Y2J0), RP3A_MOUSE (P47708), RP3A_RAT (P47709), RSG4_HUMAN (O95294), RSG4_MOUSE (Q9Z268), RSG5_HUMAN (O43374), SY61_DISOM (P24505), SY62_DISOM (P24506), SY63_DISOM (P24507), SY65_APLCA (P41823), SY65_DROME (P21521), SYT1_BOVIN (P48018), SYT1_CAEEL (P34693), SYT1_CHICK (P47191), SYT1_HUMAN (P21579), SYT1_MOUSE (P46096), SYT1_RAT (P21707), SYT2_MOUSE (P46097), SYT2_RAT (P29101), SYT3_HUMAN (Q9BQG1), SYT3_MOUSE (O35681), SYT3_RAT (P40748), SYT4_MOUSE (P40749), SYT4_RAT (P50232), SYT5_HUMAN (O00445), SYT5_MOUSE (Q9R0N5), SYT5_RAT (P47861), SYT7_HUMAN (O43581), SYT7_MOUSE (Q9R0N7), SYTA_MOUSE (Q9R0N4), SYTA_RAT (O08625), YPT7_CAEEL (P41885) Above protein IDs are from SwissProt.