Motif-based Hidden Markov Models for Multiple Sequence Alignment

William N. Grundy • Charles P. Elkan
Dept. of Computer Science & Engineering
University of California, San Diego

Abstract

- Protein families are well characterized by a collection of motifs (Sonnhammer & Kahn 1994), sometimes referred to as the “common core” (Chothia & Lesk 1986).
- These motifs can have structural and functional significance, and they may frequently be operated upon as units by diverse evolutionary mechanisms.
- The quality of a multiple alignment depends upon how accurately it identifies an ordered series of motifs (McClure et al. 1994, 1995).
- Hidden Markov models (HMMs) provide a theoretically sound modeling paradigm for collections of motifs for which efficient algorithms exist.
- Meta-MEME (Grundy et al. 1996, 1997) is a software toolkit that builds left-to-right, motif-based HMMs that focus upon the common core.
- Meta-MEME has been shown to detect remote homologies using smaller training sets than are required by standard HMMs.
- In an analysis of four protein families, Meta-MEME alignments are shown to be of higher quality than those produced by standard HMMs.
- In a previous analysis of nine other multiple alignment methods (McClure et al. 1994), only one method yields significantly higher quality alignments than Meta-MEME.

Meta-MEME

- **Step 1**: MEME builds motif models with maximal posterior probability, given a set of related sequences.
- **Step 2**: MAST uses only the motifs that appear in more than half the training sequences to search a database for the typical order and spacing of those motifs.
- **Step 3**: Meta-MEME constructs a hidden Markov model that combines the motif models in a linear fashion according to the typical schema.

Methods

- Four families were analyzed: globins, eukaryotic kinases, aspartic acid proteases, and the RH domain of the RNA-directed DNA polymerase (reverse transcriptase).
- Each family was represented by a test set of twelve sequences (McClure et al. 1994).
- Sequence identity among the sequences was 10-70% for the globins and 8-30% for all other families.
- Standard HMMs (Krogh et al. 1994) were created using the HMMER v1.8 tool `hmmt` (Eddy 1995).
- Meta-MEME and HMMER models were aligned using `hmma` with default parameter settings.

Multiple Alignments

- Pre-determined test motifs are indicated by color.
- The `x`'s above the alignments indicate Meta-MEME motif positions or HMMER match state positions.
- Amino acids are unaligned outside of motifs in Meta-MEME alignments or within insertions in HMMER alignments.

Meta-MEME Alignment of Globin Sequences

```
          ........................xxxxxxxx............xxxxxxxxxxxxxxxx.......
HAHU      vlspa.................dktnVKAAWGKVgahag.....eygaEALERMFLSFPTTKTYFphfd...
HAOR      mltda.................ekkeVTALWGKAaghge.....eygaEALERLFQAFPTTKTYFshfd...
HADK      vlsaa.................dktnVKGVFSKIgghae.....eygaETLERMFIAYPQTKTYFphfd...
HBHU      vhltp.................eeksaVTALWGKVnvde.......vggEALGRLLVVYPWTQRFFesfgdls
HBOR      vhlsg.................geksaVTNLWGKVnine.......lggEALGRLLVVYPWTQRFFeafgdls
HBDK      vhwta.................eekqlITGLWGKVnvad.......cgaEALARLLIVYPWTQRFFasfgnls
MYHU      glsdg.................ewqlVLNVWGKVeadip.....ghgqEVLIRLFKGHPETLEKFdkfkhlk
MYOR      glsdg.................ewqlVLKVWGKVegdlp.....ghgqEVLIRLFKTHPETLEKFdkfkglk
IGLOB     mkffavlalcivgaiaspltadeaslVQSSWKAVshn.........evEILAAVFAAYPDIQNKFsqf....
GPUGNI    altek.................qealLKQSWEVLkqnip.....ahslRLFALIIEAAPESKYVFsflkds.
GPYL      gvl.....................tdVQVALVKSsfeefnanipknthRFFTLVLEIAPGAKDLFsflkgs.
GGZLB     mldqqtin...........iikatvpVLKEHGVTit..........tTFYKNLFAKHPEVRPLFdmg....

          ......xxxxxxxxxxxx.............xxxxxxxxx.........xxxxxxxxxxxxxx......
HAHU      ...lshGSAQVKGHGKKVadalt......navaHVDDMPNALsa......lSDLHAHKLRVDPVNFk.....
HAOR      ...lshGSAQIKAHGKKVadals......taagHFDDMDSALsa......lSDLHAHKLRVDPVNFk.....
HADK      ...lshGSAQIKAHGKKVaaalv......eavnHVDDIAGALsk......lSDLHAQKLRVDPVNFk.....
HBHU      tpdavmGNPKVKAHGKKVlgafs......dglaHLDNLKGTFat......lSELHCDKLHVDPENFr.....
HBOR      sagavmGNPKVKAHGAKVltsfg......dalkNLDDLKGTFak......lSELHCDKLHVDPENFn.....
HBDK      sptailGNPMVRAHGKKVltsfg......davkNLDNIKNTFaq......lSELHCDKLHVDPENFr.....
MYHU      sedemkASEDLKKHGATVltalggi..lkkkghHEAEIKPLAq........SHATKHKIPVKYLEF......
MYOR      tedemkASADLKKHGGTVltalgni..lkkkgqHEAELKPLAq........SHATKHKISIKFLEY......
IGLOB     ...agkDLASIKDTGAFAt..............HATRIVSFLse......vIALSGNTSNAAAVNSlvsklg
GPUGNI    .neipeNNPKLKAHAAVIfkticesatelrqkgHAVWDNNTLkr......lGSIHLKNKITDPHFEvmk...
GPYL      .sevpqNNPDLQAHAGKVfklt.......yeaaIQLEVNGAVasdatlkslGSVHVSKGVVDAHFPvvk...
GGZLB     ...rqeSLEQPKALAMTVlaa.........aqNIENLPAILpav...kkiAVKHCQAGVAAAHYPi.....

          ............xxxxxxxxxxxxxxxxxxxxxxxxx...........xxxxxxxxxx
HAHU      ............lLSHCLLVTLAAHLPAEFTPAVHASLDKfl........asVSTVLTSKYR
HAOR      ............lLAHCILVVLARHCPGEFTPSAHAAMDKfl........skVATVLTSKYR
HADK      ............fLGHCFLVVVAIHHPAALTPEVHASLDKfm........caVGAVLTAKYR
HBHU      ............lLGNVLVCVLAHHFGKEFTPPVQAAYQKvv........agVANALAHKYH
HBOR      ............rLGNVLIVVLARHFSKDFSPEVQAAWQKlv........sgVAHALGHKYH
HBDK      ............lLGDILIIVLAAHFTKDFTPECQAAWQKlv........rvVAHALARKYH
MYHU      ............ISECIIQVLQSKHPGDFGADAQGAMNKalelf..rkdmaSNYKELGFQG
MYOR      ............ISEAIIHVLQSKHSADFGADAQAAMGKalelf..rndmaAKYKEFGFQG
IGLOB     ddhkargvsaaqfGEFRTALVAYLQANVSWGDNVAAAWNKal.........dNTFAIVVPRL
GPUGNI    ..........gaLLGTIKEAIKENWSDEMGQAWTEAYNQl..........VATIKAEMKE
GPYL      ..........eaILKTIKEVVGDKWSEELNTAWTIAYDEla.........iIIKKEMKDAA
GGZLB     ............VGQELLGAIKEVLGDAATDDILDAWGKaygviadvfiqvEADLYAQAVE
```

HMMER Alignment of Globin Sequences

```
         xxxxxxxxxxx.xxxxxxxxx.xxxxx.........xxxxxx.xxxxxxx.xxxxxxxxx.xxxxxxxxx
HAHU     V.LSPADKTN..VKAAWGKVG.AHAGE.........YGAEAL.ERMFLSF..PTTKTYFPH.FDLS.HGSA
HAOR     M.LTDAEKKE..VTALWGKAA.GHGEE.........YGAEAL.ERLFQAF..PTTKTYFSH.FDLS.HGSA
HADK     V.LSAADKTN..VKGVFSKIG.GHAEE.........YGAETL.ERMFIAY..PQTKTYFPH.FDLS.HGSA
HBHU     VHLTPEEKSA..VTALWGKVN.VDEVG..........G.EAL.GRLLVVY..PWTQRFFES.FGDL.STPD
HBOR     VHLSGGEKSA..VTNLWGKVN.INELG..........G.EAL.GRLLVVY..PWTQRFFEA.FGDL.SSAG
HBDK     VHWTAEEKQL..ITGLWGKVNvAD.CG..........A.EAL.ARLLIVY..PWTQRFFAS.FGNL.SSPT
MYHU     G.LSDGEWQL..VLNVWGKVE.ADIPG.........HGQEVL.IRLFKGH..PETLEKFDK.FKHL.KSED
MYOR     G.LSDGEWQL..VLKVWGKVE.GDLPG.........HGQEVL.IRLFKTH..PETLEKFDK.FKGL.KTED
IGLOB    M.KFFAVLALCiVGAIASPLT.ADEASlvqsswkavsHNEVEIlAAVFAAY.PDIQNKFSQFaGKDLASIKD
GPUGNI   A.LTEKQEAL..LKQSWEVLK.QNIPA.........HS.LRL.FALIIEA.APESKYVFSF.LKDSNEIPE
GPYL     GVLTDVQVAL..VKSSFEEFN.ANIPK..........N.THR.FFTLVLEiAPGAKDLFSF.LKGSSEVPQ
GGZLB    M.L.DQQTIN..IIKATVPVLkEHGVT..........ITTTF.YKNLFAK.HPEVRPLFDM.GRQ..ESLE

         xxxxx.xxxxxxxxxxxxx..xxxxxxxxxxxxxxx..xxxxxxx.xxxxxxx...xxxxxxxxxxxxxxxx
HAHU     QVKGH.GKKVADA.LTN......AVA.HVDDMPNA...LSALS.D.LHAHKL....RVDPVNF.KLLSHCLL
HAOR     QIKAH.GKKVADA.L.S......TAAGHFDDMDSA...LSALS.D.LHAHKL....RVDPVNF.KLLAHCIL
HADK     QIKAH.GKKVAAA.LVE......AVN.HVDDIAGA...LSKLS.D.LHAQKL....RVDPVNF.KFLGHCFL
HBHU     AVMGNpKVKAHGK.KVLGA..FSDGLAHLDNLKGT...FATLS.E.LHCDKL....HVDPENF.RL.LGNVL
HBOR     AVMGNpKVKAHGA.KVLTS..FGDALKNLDDLKGT...FAKLS.E.LHCDKL....HVDPENFNRL..GNVL
HBDK     AILGNpMVRAHGK.KVLTS..FGDAVKNLDNIKNT...FAQLS.E.LHCDKL....HVDPENF.RL.LGDIL
MYHU     EMKASeDLKKHGA.TVL......TALGGILKKKGHH..EAEIKPL.AQSHATK...HKIPVKYLEFISECII
MYOR     EMKASaDLKKHGG.TVL......TALGNILKKKGQH..EAELKPL.AQSHATK...HKISIKFLEYISEAII
IGLOB    T.GA...FATHATRIVSFLseVIALSGNTSNAAAV...NSLVSKL.GDDHKA....R.GVSAA.QF..GEFR
GPUGNI   NNPK...LKAHAAVIFKTI...CESATELRQKGHAVwdNNTLKRL.GSIHLK....N.KITDP.HF.EVMKG
GPYL     NNPD...LQAHAG.KVFKL..TYEAAIQLEVNGAVAs.DATLKSL.GSVHVS....K.GVVDA.HF.PVVKE
GGZLB    Q......PKALAM.TVL......AAAQNIENLPAIL..PAVKKIAvKHCQAGVaaaH.YPIVGQEL.LGAIK

         xxxxxxxxx.xxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxx..x
HAHU     VT.LAA.H..LPAEFTPA..VHASLDKFLASV.STVLTS..KY..R
HAOR     VV.LAR.H..CPGEFTPS..AHAAMDKFLSKV.ATVLTS..KY..R
HADK     VV.VAI.H..HPAALTPE..VHASLDKFMCAV.GAVLTA..KY..R
HBHU     VCVLAH.H..FGKEFTPP..VQAAYQKVVAGV.ANALAH..KY..H
HBOR     IVVLAR.H..FSKDFSPE..VQAAWQKLVSGV.AHALGH..KY..H
HBDK     IIVLAA.H..FTKDFTPE..CQAAWQKLVRVV.AHALAR..KY..H
MYHU     QV.LQSKHPgDFGADAQGA.MNKALELFRKDM.ASNYKELGFQ..G
MYOR     HV.LQSKHSaDFGADAQAA.MGKALELFRNDM.AAKYKEFGFQ..G
IGLOB    TA.LVA.Y..LQANVSWGDnVAAAWNKA.LDN.TFAIVV..PR..L
GPUGNI   ALLGTIKEA.IKENWSDE..MGQAWTEAYNQLVATIKAE..MK..E
GPYL     AILKTIKEV.VGDKWSEE..LNTAWTIAYDELAIIIKKE..MKdaA
GGZLB    EVLGDAAT..DDILDAWGK.AYGVIADVFIQVEADLYAQ..AV..E
```

Meta-MEME Alignment of RH Sequences

```
          ............xxxxxxxx.........................................xxxxxxxx
HTLV-II   ldt.........apCLFSDGSPqkaayvlwdqtilqqd..........itplpshethsaqkgELLALICG
SRV-I     lnn.........alLVFTDGSStgmaaytladtti...............kfqtnlnsaqlvELQALIAV
RSV       pvp.........gpTVFTDASSsthkgvvvwregprw...........eikeiadlgasvqqlEARAVAMA
HIV-II    ipg.........aeTFYTDGSCnrqskegkagyvtdrg..........kdkvkkleqttnqqaELEAFAMA
MoMLV     pda.........dhTWYTDGSSllqegqrkagaavttet.......eviwakaldagtsaqraELIALTQA
Ingi      pre.........hyKLWTDGSVslgeklgaaallhrnntl......icapktgagelscsyraECVALEIG
CAMV      pee.........klIIETDASDdywggmlkaikinegtntelicryasgsfkaaeknyhsndkETLAVINT
17.6      ftk.........kfTLTTDASDvalgavlsqdghplsyi........srtlneheinystiekELLAIVWA
MAUP      fnnstnlqepsdsrLLYRKGSWvnirfaay.........................lysklseEKHGLVPK
HBV       rpg.........lcQVFADATPtgwglvmghqrmr..............gtfsaplpihtaELLAACFA
Copia     fen.........kiIGYVDSDWagseidrksttgylfkmfdf.nlicwntkrqnsvaassteaEYMALFEA
E.coli    mlk.........qvEIFTDGSClgnpgpggygailryrg.......rektfsagytrttnnrmELMAAIVA

          x...................xxxxxxxxxx.....................................
HTLV-II   Lraak............pwpsLNIFLDSKYLikylhslaigaflgtsahqtlqaalp.............
SRV-I     Lsafp............nqpLNIYTDSAYLahsiplletvaqikhisetaklflqcqq...........
RSV       Lllwp............ttpTNVVTDSAFVakmllkmgqegvpstaaafiledal.............
HIV-II    Ltds.............gpkVNIIVDSQYVmgisasqpteseskivnqiie.................
MoMLV     Lkmae............gkkLNVYTDSRYAfatahihgeiyrrrglltsegkeiknkdeil.........
Ingi      Lqrllkwlp....ryrstpsrLSIFSDSLSMltalqtgplavtdpilrrlwrll..............
CAMV      Ikkfsiy.........ltpvhFLIRTDNTHFksfvnlnykgdsklgrnir..................
17.6      Tktfrhy.........llgrhFEISSDHQPLswlyrmkdpnskltrwr.....................
MAUP      Flek............lreINFALDKVDVteidsklsrlmkfsvsaaydevgtlalkslfkfrnseres
HBV       Rsrs............gaNIIGTDNSVVlsrkytsfpwllgcaanwilrgtsfvyvpsa.........
Copia     Vrealwlkfllltsiniklenpikiyednqgcisiannpschkrakhidiky.................
E.coli    Lealk............ehceVILSTDSQYVrqgitqwihnwkkrgwktadkkpvknvd...........

          ........................................xxxxxxxxx.................
HTLV-II   ............pllqgktiylhhvrshtnlpdpistfNEYTDSLILapl..............
SRV-I     ............liynrsipfyighvrahsglpgpiahgNQKADLATKtvasn............
RSV       ............sqrsamaavlhvrshsevpgffteqNDVADSQATfqay.............
HIV-II    ..................emikkeaiyvawvpahkgiggNQEVDHLVSqgirqvl..........
MoMLV     .........allkalflpkrlsiihcpghqkghsaeargNRMADQAARkaaitetpdtstll...
Ingi      ..............lqvqrrkirirlqfvfdhcgvkrNEVCDEMAKkaadlpql.........
CAMV      ....................wqawlshysfdvehikgtdNHFADFLSRefnkvns..........
17.6      ....................vklsefdfdikyikgkeNCVADALSRikleety...........
MAUP      ikasfkqlrengkiaefsearrlwfeilklirldlfnasSLACDDLLShlqdrrsi.........
HBV       .........lnpaddpsrgrlglsrpllrlpfrpttgrtSLYADSPSVpshlpdrvh........
Copia     ..................hfareqvqnnvicleyiptеNQLADIFTKplpaarfve........
E.coli    ............lwqrldaalgqhqikwewvkghaghpeNERCDELARaaamnptledtgyqvev
```

HMMER Alignment of RH Sequences

```
#=RF      xxxxxxxxxxxxxx.xxxxxxx..xxxxx.xxxx..x...xxxx........xxxxxxxx....xxxxx..x
HTLV-II   LDTAPCLFSDGSPQ.KAAYVLW..DQTIL.QQDI..T...PLP.........SH.ETHS.......AQ...K
SRV-I     LNNALLVFTDGSS....TGMAA....YTL.A..D..T...TIKF........QT.NLNS.......AQ...L
RSV       PVPGPTVFTDASSS.THKGVVV....W.R.E.GP..R...WEIK........EIADLGA.....SVQQ....
HIV-II    IPGAETFYTDGSCN.RQSKEGK..AGY.V.T.DR..G...KDKV........KKLEQTT......NQQ....
MoMLV     PDADHTWYTDGSSL.LQEGQRK..AGAAV.T.TE..T...EVIW........AK.ALDAG....TSAQ...R
Ingi      PREHYKLWTDGSVS.LGEKLGA...AALL.H.RN..N...TLIC........AP.KTGA.....GELSCsyR
CAMV      PEEKLIIETDASDD.YWGGMLK..AIKIN.EGTN..T...ELICryasgsfkAA.EKNY.....HSND....
17.6      FTKKFTLTTDASDVaLGAVLSQ..DGHPL.S.YIs.R...TLN.........EH.EINY......STI...E
MAUP      FNNSTNLQEP.SDS.RLLYRKG...SWVN.I.RFaaY...LYSK........LSEEKHGLvpkfLEKL...R
HBV       RPGLCQVFADATP..TGWGLVM...GHQR.M..R..G...TFSA........PL.PIHT.....AELL....
Copia     FENKIIGYVDSDWA.GSEIDRKstTGYLFkMFDF..N...LICW........NTKRQNSVa...ASST...E
E.coli    MLKQVEIFTDGSCL.GNPGPGG..YGAIL.R.YR..GrekTFSA........GYTRTTNNr...MELM....

#=RF      xxxxxxx......xxxxxx.xxxxxxx.xxx.xxxxxxx..xxxx.xxxxxxxxxxxx....xxxxxxxx..
HTLV-II   GELLAL.......ICGLRAaKPWPSL..NI..FLD.S....KYLI.KYL.HSLAIGAFl...GTSAHQ....
SRV-I     VELQAL.......IAVLSA.FPNQPL..NI..YTD.S....AYLA.HSI.PLLETVAQikhiSETAKL....
RSV       LEARAV.......AMAL...LLWPTTP.TNV.VTDSAF...VAKM.LLK.MGQEGV.P....STAAAF....
HIV-II    AELEAF.......AMAL...TDSGPKV.NI..IVD.S....QYVM.GIS.ASQPTESE....SKIVNQ....
MoMLV     AELIALTqa...lKMAEGK.KLNVYTD.SRYaFATAHI...HGEI.YRR.RGLLTS.E....GKEIKN....
Ingi      AECVALE......IGLQRL.LKWLPRYrST..PSRLSIFs.DSLS.MLT.ALQTGPLA....VTDPIL....
CAMV      KETLAV.......INTI.K.KFSIYL..TPVhFLI.R....TDNT.HFK.SFVNLN.Y....KGDSKL....
17.6      KELLAI.......VWAT.K.TFRHYLL.GRH.FEISSD...HQPL.SWL.YR.M.KD......PNSKL....
MAUP      EINFALDkvdvteIDSKLS.RLMKFSV.SA..AYDEVGTl.ALKS.LFKFRNSERESI....KASFKQLRen
HBV       AACFARSr.....SGANIIgTDNSVVL.SR..KYTSFPWllGCAA.NWILRGTSFV.Y....VPSALNPA..
Copia     AEYMAL.......FEAVREaLWLKFLL.TS..INIKL....ENPI.KIYEDNQGCISIa..nNPSCHKRAk.
E.coli    AAIVALE......ALKEHC.EVILSTD.SQ..YVRQGI...TQWIhNWKKRGWKTADK....KPVKNV....

#=RF      .....xxxxxxxxxxxxxxxx.xxxxxx.xxxxxxxxx..xxxxxxxxxxxx...xxxx.xxx...xxxx
HTLV-II   .....TLQA..ALPPLLQGKT....IYLH.HVRSHT......N.LPDPISTFN...EYTDsLIL...APL.
SRV-I     .....FLQCQ.QLIYNRSIPF....YIGH.V.RAHS......G.LPGPIAHGN...QKAD.LATktvASN.
RSV       .....ILED..ALS..QRSAM...AAVLH.V.RSHSEVPgfFT.EGNDVADS....QAT..F.....QAY.
HIV-II    .....IIEEM.IKK..EAIYV...AWVPA.H.KGIGG.....NQEV.D.HLV....SQG...IR...QVL.
MoMLV     .....KDEIL.ALLKALFLPKRlSIIHCPgHQKGHSAEAr.GN.RMADQAARKaaiTET..PDT...STLL
Ingi      .....RRLWR.LLLQVQRRKI...RIRLQ.FVFDHCGVK..RN.EVCD.EMA....KKA..ADL...PQL.
CAMV      .....GRNIRWQAW..LSHYS...FDVEH.I.KGTD......N.HFAD.FLS....REF..N.....KVNS
17.6      .....TRWRV.KLS..EFDFD....IKY..I.KGKE......N.CVAD.ALSR...IKL..E.....ETY.
MAUP      gkiaeFSEAR.RLW.FEILKL....IRLD.L.FNAS......S.LACD.DLL....SHL..QDR...RSI.
HBV       .....DDPSRGRLG..LSRPL....LRLP.F.RPTTGR....TSLYAD.SPSV...PSHL.PD....RVH.
Copia     ....hIDIKYHFAR.EQVQNN...VICLE.Y.IPTE......N.QLADIFTK....PLP..AAR...FVE.
E.coli    .....DLWQRLDAALGQHQIKW.EWVKGH.AGHPENER....CDELAR.AAAM...NPTL.EDTg.yQVEV
```

Alignment Scores

Globins	1	2	3	4	5	Total
META-MEME	12	11	11*	11*'	12*	57
HMMER	11	12*'	11*	10*'	11*'	55
AMULT	12	12	12	12	12	60
ASSEMBLE	12	11	12	12	12	59
CLUSTAL V	12	11	12	12	12	59
DFALIGN	12	12	12	12	12	60
GENALIGN	11*'	12	12	10'	11	56
MULTAL	12	11	12	12	12	59
MACAW	9	11	9	8	8	45
PIMA	12	12	12	12	12	60
PRALIGN	8	8*	9*	8*	10	43

Kinases	1	2	3	4	5	6	7	8	Total
META-MEME	12	11	10	12	12	12	12	11	92
HMMER	12	12*	12*	12	12	12	12	12	96
AMULT	12	10	11	12	12	12	12	12	93
ASSEMBLE	11	7	10	12	12	12	12	12*	87
CLUSTALV	12	11	11*	12	12	12	12	12*	94
DFALIGN	12	12	12	12	12	12	12	12	96
GENALIGN	12'	9*	10	12	12	12	12*	11*	90
MULTAL	12	9*	10*	12	12	12*	12	12	91
MACAW	8	0	9	12	12	10	12	0	63
PIMA	12	11	11	12	12	12	12	12	94
PRALIGN	12	10*	6*	4	9*	9*	4	4	58

Proteases	1	2	3	Total
META-MEME	12	4*'	9*	25
HMMER	11	8	6*	25
AMULT	11	7	10	28
ASSEMBLE	—	—	—	0
CLUSTALV	12	9*	6*	27
DFALIGN	12	12*	12	36
GENALIGN	11	8*'	7*	26
MULTAL	10	7*	9*	26
MACAW	12	4	8	24
PIMA	12	5*	5*	22
PRALIGN	8*	4*	8*	20

RHs	1	2	3	4	Total
META-MEME	11	9	11	12	43
HMMER	11	10*	5*'	7*	33
AMULT	11	9*	8*	7*	35
ASSEMBLE	—	—	—	—	0
CLUSTALV	12	9	9*	9*	39
DFALIGN	12	12	10	12	46
GENALIGN	12*'	7	8*'	9*'	36
MULTAL	11*	11*	9*	10	41
MACAW	7	5	7	3	22
PIMA	10	9	8*	11*	38
PRALIGN	9	8*	6*	3	26

Scores represent the number of sequences in which each motif was correctly aligned. Scores for non-HMM methods are from (McClure et al. 1994). A * indicates that the motif was correctly aligned in two or more misaligned subsets of the test sequences. A ' indicates that a gap was inserted into the motif.

Results

	Globins	Kinases	Proteases	RHs	Total
DFALIGN	60	96	36	46	238
CLUSTALV	59	94	27	39	219
META-MEME	**57**	**92**	**25**	**43**	**217**
MULTAL	59	91	26	41	217
AMULT	60	93	28	35	216
PIMA	60	94	22	38	214
HMMER	**55**	**96**	**25**	**33**	**209**
GENALIGN	56	90	26	36	208
MACAW	45	63	24	22	154
PRALIGN	43	58	20	26	147
ASSEMBLE	59	87	0	0	146

- Meta-MEME ranks third overall.
- MEME discovers 19 of the 20 motifs in the test set.
- Only DFALIGN (Feng & Doolittle, 1987) significantly outperforms other methods.

Discussion

- Meta-MEME's focus on the common core allows accurate models to be trained from fewer sequences than are required by standard HMMs.
- By focusing its models on highly conserved regions of the training set, Meta-MEME effectively ignores noisy portions of the data, thereby allowing the software to properly align distant homologs.
- Meta-MEME alignments could be extended by running another alignment program on the unaligned, non-motif regions.
- A server for MEME and MAST, as well as the source code for Meta-MEME, is available at **`http://www.sdsc.edu/MEME`**.

References

Bailey, T. L. and C. P. Elkan. "Fitting a mixture model by expectation-maximization to discover motifs in biopolymers." *2nd ISMB*, 1994.

Chothia, C. and A. M. Lesk. "The relation between the divergence of sequence and structure in proteins." *EMBO*, 5:823-826, 1986.

Eddy, S. R. "Multiple Alignment Using Hidden Markov Models." *3rd ISMB*, 114-120, 1995.

Feng, D.-F. and R. F. Doolittle. "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." *J. Mol. Evol.*, 25:351-360, 1987.

Grundy, W. N., T. L. Bailey, C. P. Elkan and M. E. Baker. "Meta-MEME: Motif-based hidden Markov models of protein families." *CABIOS*, 1997. To appear.

Grundy, W. N., T. L. Bailey, C. P. Elkan and M. E. Baker. "Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs." *BBRC*, 231(3):760-766, 1997.

Krogh, A., M. Brown, I. Mian, K. Sjolander and D. Haussler. "Hidden Markov Models in Computational Biology: Applications to Protein Modeling." *JMB*, 234:1501-1531, 1994.

McClure, M. A., T. K. Vasi and W. M. Fitch. "Comparative analysis of multiple protein-sequence alignment methods." *Mol. Bio. Evol.*, 11(4):571-592, 1994.

McClure, M. A. and R. Raman. "Parametrization studies of hidden Markov models representing highly divergent protein sequences." *28th HICSS*, 184-193, 1995.

Sonnhammer, E. and D. Kahn. "The modular arrangement of proteins as inferred from analysis of homology." *Protein Science*, 3:482-492, 1994.