Biochemical and Biophysical Research Communications 231(6):760–766, 1997.

Hidden Markov Model Analysis of Motifs in Steroid Dehydrogenases and their Homologs

William N. Grundy bgrundy@cs.ucsd.edu
Department of Computer Science and Engineering
University of California, San Diego (619) 453-4364
La Jolla, California 92093-0114 FAX (619) 534-7029

Timothy L. Bailey tbailey@sdsc.edu
San Diego Supercomputer Center
P.O. Box 85608 (619) 534-8350
San Diego, California 92186-9784 FAX (619) 534-5127

Charles P. Elkan elkan@cs.ucsd.edu
Department of Computer Science and Engineering
University of California, San Diego (619) 534-8897
La Jolla, California 92093-0114 FAX (619) 534-7029

Michael E. Baker[*] mbaker@ucsd.edu
Department of Medicine
University of California, San Diego (619) 534-4164
La Jolla, CA 92093-0623 FAX (619) 534-1424

[*]Corresponding author.

Abstract

The increasing size of protein sequence databases is straining methods of sequence analysis, even as the increased information offers opportunities for sophisticated analyses of protein structure, function and evolution. Here we describe a method called Meta-MEME that uses artificial intelligence-based algorithms to build models of families of protein sequences. These models can be used to search protein sequence databases for remote homologs. The MEME (Multiple Expectation-maximization for Motif Elicitation) software package identifies motif patterns in a protein family, and these motifs are combined into a hidden Markov model (HMM) that can be used as a database searching tool. Meta-MEME is sensitive and accurate, as well as automated and unbiased, making it suitable for the analysis of large datasets. We demonstrate Meta-MEME on a family of dehydrogenases that includes mammalian 11β-hydroxysteroid and 17β-hydroxysteroid dehydrogenase and their homologs in the short chain alcohol dehydrogenase family. We chose this dataset because it is large and phylogenetically diverse, providing a good test of the sensitivity and selectivity of Meta-MEME on a protein family of biological interest. Indeed, Meta-MEME identifies at least 350 members of this family in Genpept96 and clearly separates these sequences from non-homologous proteins. We also show how the MEME motif output can be used for phylogenetic analysis.

1 Introduction

The number of known protein sequences is increasing rapidly as various genome projects come on line [1, 2, 3]. This explosion of data provides an opportunity for comparisons of protein sequences from distantly related organisms, allowing for the identification of conserved regions, or motifs, that are likely to be functionally important. The usual approach for identifying distantly related homologs is to search a database with a sequence using FASTA [4] or BLAST [5]. However, as databases increase in size, such searches tend to miss the more distantly related homologs because of the noise from unrelated proteins having a random similarity to the sequence being searched.

Sensitivity can be increased by using the information from several homologous proteins to construct a composite of conserved regions for database searching [6, 7, 8]. In this approach, homologous sequences are aligned, conserved motifs are identified and an amino acid profile or log-odds matrix for each motif is calculated. This log-odds matrix is representative of the relative amino acid probabilities at specific positions and is characteristic of the protein family, which makes the log-odds matrix a sensitive probe for searching a database. Increasing the number of diverse protein sequences for motif analysis increases the sensitivity of the resulting search, as well as increasing the information about motif structure and its relationship to function. Unfortunately, aligning a large number of divergent sequences requires gaps and insertions. These complicate the multiple sequence alignment, and in some cases, make it difficult to accurately characterize the boundaries of the motifs, reducing their utility for analysis of structure and function.

To address these problems, we have developed an automated method for constructing motifs called Multiple Expectation-maximization for Motif Elicitation (MEME) [9, 10]. This method can analyze large datasets — in this paper we use thirty-seven dehydrogenase sequences — using a statistical algorithm called expectation-maximization [11]. MEME discovers a set of motifs that describe the given group of related sequences. The unbiased and automated properties of this method make it accurate and convenient for determining motifs. Moreover, each motif's log-odds matrix is a sensitive probe for searching a databank such as Genpept96 or SWISSPROT for distantly related homologs. A version of MEME running on a parallel supercomputer is available via the World-Wide Web at `http://www.sdsc.edu/MEME`.

Here we describe improvements that increase the sensitivity and selectivity of this method by incorporating into the searching algorithm two other important pieces of information from the motif analysis: the order and spacing of motifs. To use this information, we have created Meta-MEME, an automated hidden Markov model extension to MEME. Hidden Markov models have been used previously to characterize protein families and to direct homology searches [12, 13]. Meta-MEME differs from these other HMM

approaches in its focus on motif regions. By precisely modeling only the highly-conserved regions of the dataset, Meta-MEME selectively discards noisy, inter-motif information.

In addition, we show that concatenated MEME motifs can be used to construct reliable phylogenetic trees for distantly related sequences. Concatenated motifs can be aligned unambiguously, unlike entire sequences. This is an important consideration when constructing a multiple alignment of many distantly related sequences because the alignment may be degraded by mutations suggested spuriously by ambiguities in assigning insertions and deletions. Others have dealt with this problem and have constructed useful phylogenetic trees by ignoring the highly divergent segments containing insertions and deletions [14, 15]. We find that concatenated MEME motifs also yield useful trees, with the advantage that the analysis is unbiased and automated.

We use Meta-MEME to analyze a family of dehydrogenases [16, 17, 18, 19, 20, 21] that includes 11β-hydroxysteroid and 17β-hydroxysteroid dehydrogenase, enzymes that are important in actions of steroids that affect blood pressure, reproduction and development and also the growth of some cancers of breast and prostate. In addition to its medical importance, we chose this family for testing our method because it is large and phylogenetically diverse and, thus, representative of what will be available for analysis as more genomes are sequenced. Using a dataset of thirty-seven dehydrogenases, we find that Meta-MEME is a sensitive, selective and convenient tool for identifying distantly related homologs in databases, which should prove useful for subsequent analysis of their structure, function and evolution.

2 Methods

The details of the MEME algorithm have been described elsewhere [9, 10]. Briefly, MEME uses the expectation-maximization algorithm [11] to disover conserved regions, or motifs, in a dataset of protein sequences. The algorithm uses a heuristic criterion function based on a maximum likelihood ratio test to compare candidate motifs. MEME outputs models of conserved regions in rank order, with the strongest motif represented by the first model. For the analyses reported here, we use MEME version 2.0 with the minimum width set at 12 amino acids and the Dirichlet mixture prior [9, 10]. The training set consists of the thirty-seven sequences shown in Table 1 with their SWISSPROT codes. Pairwise alignments of almost all of these sequences are less than 30% identical after using gaps and insertions to maximize identities [17, 22, 23]. Many sequences are less than 20% identical after use of gaps and insertions.

Hidden Markov models

A hidden Markov model is a probabilistic model in which a hidden stochastic process produces a sequence of observable outputs [24]. In Meta-MEME, the sequence of outputs is a series of amino acids. The model is linear, and each hidden state in the model corresponds to one or more adjacent amino acids in the protein family being modeled. In a Meta-MEME hidden Markov model, motif regions are modeled without insert states, so the motifs are similar to gapless profiles [6]. Spacer regions between motifs can be of variable length.

The six strongest motifs in the set of thirty-seven divergent dehydrogenase sequences are determined using MEME 2.0. Then Genpept96 is searched with the log-odds output for the six motifs, and the highest scoring protein is used as a canonical template for the motif order and spacing. This template provides the framework for a motif-based hidden Markov model incorporating all six motifs. Because the hidden Markov model is linear, it takes into account the canonical order and spacing of the motifs. The motif-based hidden Markov model is used by a modified Smith-Waterman algorithm [25] to search Genpept96 for homologs. The output score for each sequence is expressed in bits (i.e., log_2).

2BHD_STREX	20-Beta-Hydroxysteroid Dehydrogenase
3BHD_COMTE	3-Beta-Hydroxysteroid Dehydrogenase
ACT3_STRCO	Putative Ketoacyl Reductase
ADH_DROME	Alcohol Dehydrogenase
AP27_MOUSE	Adipocyte P27 Protein (AP27).
BA72_EUBSP	7-Alpha-Hydroxysteroid Dehydrogenase
BDH_HUMAN	D-Beta-Hydroxybutyrate Dehydrogenase Precursor
BEND_ACICA	Cis-1,2-Dihydroxy-3,4-Cyclohexadiene-1-Carboxylate Dehydrogenase
BPHB_PSEPS	Biphenyl-2,3-Dihydro-2,3-Diol Dehydrogenase
BUDC_KLETE	Acetoin(Diacetyl) Reductase
CSGA_MYXXA	C-Factor.
DHB2_HUMAN	Estradiol 17 Beta-Dehydrogenase 2
DHB3_HUMAN	Estradiol 17 Beta-Dehydrogenase 3
DHCA_HUMAN	Carbonyl Reductase (NADPH)
DHES_HUMAN	Estradiol 17 Beta-Dehydrogenase
DHGB_BACME	Glucose 1-Dehydrogenase B
DHII_HUMAN	Corticosteroid 11-Beta-Dehydrogenase
DHMA_FLAS1	N-Acylmannosamine 1-Dehydrogenase
ENTA_ECOLI	2,3-Dihydro-2,3-Dihydroxybenzoate Dehydrogenase
FABG_ECOLI	3-Oxoacyl-[Acyl-Carrier Protein] Reductase
FABI_ECOLI	Enoyl-[Acyl-Carrier-Protein] Reductase (NADH)
FIXR_BRAJA	FixR Protein.
FVT1_HUMAN	Follicular Variant Translocation Protein 1 Precursor (FVT-1).
GUTD_ECOLI	Sorbitol-6-Phosphate 2-Dehydrogenase
HDE_CANTR	Hydratase-Dehydrogenase-Epimerase (HDE).
HDHA_ECOLI	7-Alpha-Hydroxysteroid Dehydrogenase
HMTR_LEIMA	H Region Methotrexate Resistance Protein
LIGD_PSEPA	C Alpha-Dehydrogenase
MAS1_AGRRA	Agropine Synthesis Reductase.
NODG_RHIME	Nodulation Protein G (Host-Specificity Of Nodulation Protein C).
PCR_PEA	Protochorophyllide Reductase Precursor
PGDH_HUMAN	15-Hydroxyprostaglandin Dehydrogenase (NAD(+))
PHBB_ZOORA	Acetoacetyl-Coa Reductase
RIDH_KLEAE	Ribitol 2-Dehydrogenase
YINL_LISMO	Hypothetical 26.8 Kd Protein In Inla 5'region (ORFA).
YRTP_BACSU	Hypothetical 25.3 Kd Protein In Rtp 5'region (ORF238)
YURA_MYXXA	Hypothetical Protein In Uraa 5'region (Fragment).

Table 1: SWISSPROT identifiers and descriptions for the 37 short chain alcohol dehydrogenase training set.

Phylogeny

The sequences of the first six motifs from the MEME analysis of each dehydrogenase homolog were collapsed into a single string. These motif-only strings were analyzed using the protein parsimony analysis program from the Phylip software package [26]. The analysis was repeated 30 times, using at each iteration a random reordering of the sequences, and selecting the most parsimonious tree from all iterations.

3 Results

MEME analysis

Figure 1 displays the six motifs of the dehydrogenase dataset along with the entropy plot, which is a measure of the information content at each position. The motifs are mapped onto the primary sequence of 20β-hydroxysteroid dehydrogenase in Figure 2. Also shown in Figure 2 is the secondary structure determined from X-ray crystallographic analysis [27]. The secondary and tertiary structure of this enzyme is very similar to homologs such as dihydropteridine reductase [28], 17β-hydroxysteroid dehydrogenase-type 1 [29], enoyl reductases [30, 31], and *E. coli* 7α-hydroxysteroid dehydrogenase [32] despite having pairwise sequence similarities of 15% to 22%. The six motifs map onto structurally important domains, some of which have been shown to be functionally important by site-specific mutagenesis studies [33, 34, 35, 36, 37, 38] and structural analysis [39, 40]. Beginning at the amino terminus, the order of the motifs is (2)-(1)-(6)-(4)-(3)-(5). Their combined length is 85 amino acids, and they span 183 residues of 20β-hydroxysteroid dehydrogenase.

Hidden Markov model analysis

These six motifs were combined in their proper order into a single hidden Markov model for analysis. This model was then used to search Genpept96 for homologs. Figure 3 shows the histogram of the output of this search, and Table 2 shows selected sequences from the output. The distribution is bimodal with a clear minimum at 20 bits, demonstrating excellent separation of dehydrogenase homologs from the rest of the database.

The high scoring sequences contain the full 85 residues in the template, which spans 180 to 188 amino acids in most of the proteins. This is consistent with an absence of extra loops in these proteins and a common 3D structure. An interesting exception is carbonyl reductase, in which the six motifs span 228 residues due to an insertion of 41 residues between motifs 4 and 2 [41]. This insertion does not compromise the analysis. Meta-MEME output is useful in identifying the region where a distantly homologous protein has diverged from the dataset. For example, *Drosophila immigrans* alcohol dehydrogenase has a score of 90.6 bits based on residues 14-85 of the template. Evidently, the segment corresponding to motif 2 in this alcohol dehydrogenase has diverged from the dataset. A similar analysis holds for an oxidoreductase (score of 65.7 bits) required for shoot apex development in *Arabidopsis thaliana*.

We examined the sequences with scores below twenty bits using citations in Entrez and SwissProt and, in some cases, a BLAST search to determine which sequences were homologous to short chain dehydrogenases. All sequences above 8.9 bits are homologs. The first non-homologous protein is malate dehydrogenase at 8.9 bits; the next is ribulose bisphosphate carboxylase/oxygenase at 8.5 bits.

Phylogeny

One consequence of the projects to sequence genomes in phylogenetically diverse organisms is a wider use of phylogenetic analysis to assist in understanding the evolution of structure and function. We were interested in how well the motifs generated by MEME could be used for a phylogenetic analysis. We therefore

```
MOTIF 1
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1
content      3.4         *
(26.1 bits)  2.7    *   **
             2.0 *  * * ** *
             1.4 * ** ******
             0.7 ***********
             0.0 -----------

Multilevel       GRVDVLVNNAG
consensus          L I I
sequence
```

```
MOTIF 4
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1
content      3.4
(18.2 bits)  2.7        *
             2.0 * ***  *
             1.4 * ******
             0.7 ********
             0.0 --------

Multilevel       GRIVNVSS
consensus           I I
sequence
```

```
MOTIF 2
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1
content      3.4
(25.1 bits)  2.7    **     *
             2.0   ***   ***
             1.4 ******  ***
             0.7 ***********
             0.0 -----------

Multilevel       ALVTGASSGIG
consensus        VII  GA  L
sequence              G
```

```
MOTIF 5
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1
content      3.4
(28.1 bits)  2.7         *
             2.0  *      **   *
             1.4 ** ** * **   * *
             0.7 ***** **** * ***
             0.0 ----------------

Multilevel       GIRVNAVxPGxVxTDM
consensus         VT   I    I
sequence
```

```
MOTIF 3
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1 *
content      3.4 *   *
(31.3 bits)  2.7 *   *
             2.0 *  **     *
             1.4 * *****   **  **
             0.7 ******** *******
             0.0 ----------------

Multilevel       YSASKAAVxGLTRSLA
consensus           A FGL  FSK
sequence
```

```
MOTIF 6
        bits 6.8
             6.1
             5.4
             4.7
Information  4.1
content      3.4        *
(35.7 bits)  2.7        *
             2.0 *      *
             1.4 *     **  * *         *
             0.7 ********* ******* *** *
             0.0 -----------------------

Multilevel       WDRVIxVNLTGVFxLTRAVLPxM
consensus        F     I   S V G Q     L
sequence
```

Figure 1: **Motifs from MEME analysis of short chain alcohol dehydrogenases.** The entropy plot is a measure of the information content at each position of the motif. The consensus sequence below the entropy plot shows sites where specific amino acids are present with a probability of at least 20%.

```
                 β-strand-A  Turn         α-helix-B
                ├──────────┼────┼──────────────────────┤
  1 M N D L S G K T V I I T G G A R G L G A E A A R Q A V A A
                          2 2 2 2 2 2 2 2 2 2 2

      β-strand-B            α-helix-C
    ├──────────────┼──────────────────────┤         ├─────
 30 G A R V V L A D V L D E E G A A T A R E L G D A A R Y Q H

    β-strand-C          α-helix-D                β-strand-D
    ─────────┤ ├──────────────────────┤   ├──────────────
 59 L D V T I E E D W Q R V V A Y A R E E F G S V D G L V N N
                                            1 1 1 1 1 1 1 1 1

     β-strand-D                      α-helix-E
    ──────────┤            ├──────────────────────────────
 88 A G I S T G M F L E T E S V E R F R K V V D I N L T G V F
    1 1                           6 6 6 6 6 6 6 6 6 6 6 6 6

        α-helix-E                  β-strand-E
    ──────────────────┤          ├──────────────┤
117 I G M K T V I P A M K D A G G G S I V N I S S A A G L M G
    6 6 6 6 6 6 6 6 6 6            4 4 4 4 4 4 4 4

                        α-helix-F
            ├──────────────────────────────────┤
146 L A L T S S Y G A S K W G V R G L S K L A A V E L G T D R
                3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3            5

         β-strand-F
    ├──────────────────┤
175 I R V N S V H P G M T Y T P M T A E T G I R Q G E G N Y P
    5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

                              α-helix-G
                        ├──────────────────┤
204 N T P M G R V G N E P G E I A G A V V K L L S D T S S Y V

       β-strand-G
     ├──────────┤
233 T G A E L A V D G G W T T G P T V K Y V M G Q 255
```

Figure 2: **Alignment of MEME motifs on** *Streptomyces hydrogenans* **20β-hydroxysteroid dehydrogenase.** Each motif as determined by MEME is shown below the sequence of *S. hydrogenans* 20β-hydroxysteroid dehydrogenase. The secondary structure was determined from the X-ray analysis of crystals of *S. hydrogenans* 20β-hydroxysteroid dehydrogenase [27], and has a similar fold to that of its homologs [28, 29, 30, 31, 32]. The boxed segment at the beginning of motif 3 contains the conserved tyrosine and lysine residues at the catalytic site.

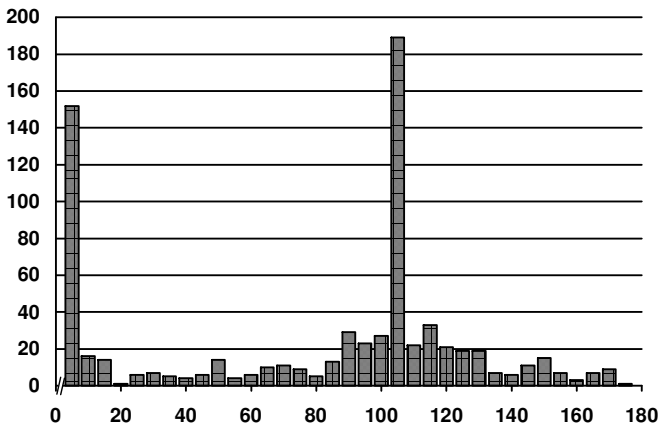

Figure 3: **Hidden Markov model analysis of Genpept96.** The output histogram has a minimum at 20 bits, demonstrating the selectivity of the HMM analysis. Sequences with negative scores are not shown. The peaks at 105 and 115 bits are due to *Drosophila* alcohol dehydrogenase sequences.

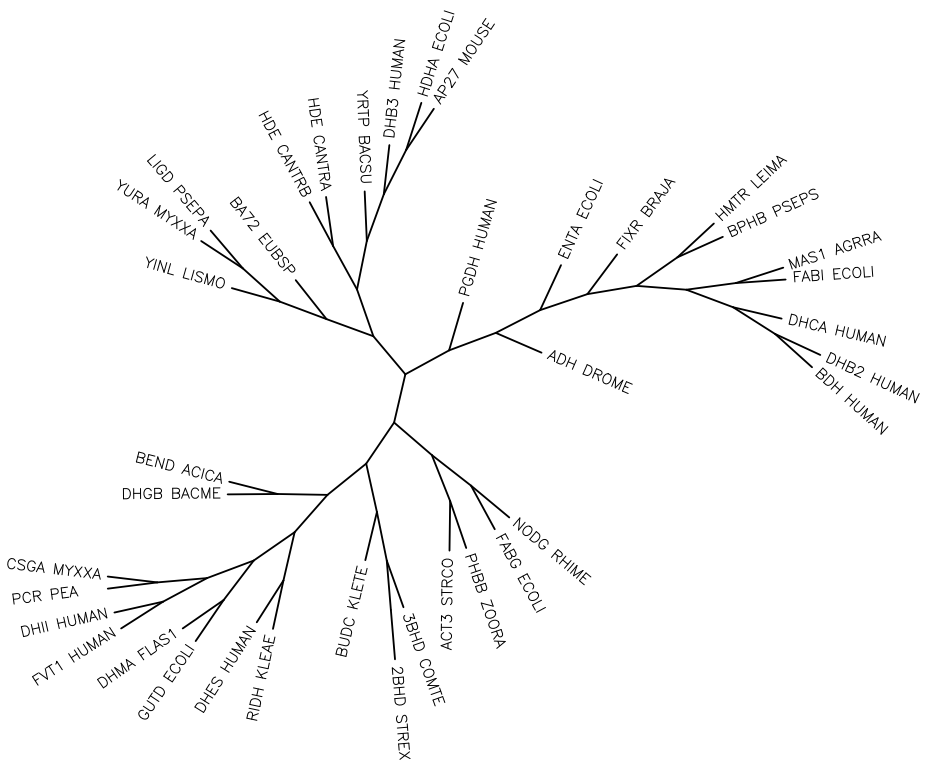

Figure 4: **Phylogenetic analysis of the dehydrogenase dataset.** The sequences of the first six motifs from the MEME analysis of each protein were collapsed into a single sequence and analyzed by parsimony analysis [26]. The 11β-hydroxysteroid and 17β-hydroxysteroid dehydrogenases-type 1 cluster together on a branch separate from 17β-hydroxysteroid dehdyrogenases-type 2 and 3, which are on separate branches. The motif phylogeny is in agreement with a phylogenetic analysis of the entire sequences of the steroid dehydrogenases [23].

Score	Sequence	Model	ID	Description
178.7	8-188	1-85	gi—145881	3-ketoacyl-acyl carrier protein reductase [Escherichia coli]
174.8	9-194	1-85	gi—153142	actIII protein [Streptomyces coelicolor]
170.9	5-184	1-85	gi—790552	acetoacetyl CoA reductase [Rhizobium meliloti]
170.4	8-190	1-85	gi—1203984	NAD+-dependent 15-hydroxyprostaglandin dehydrogenase [H. sapiens]
170.2	9-189	1-85	gi—46308	nodG gene product (AA 1-245) [R. meliloti] ketoreductase [S. nogalater]
169.3	6-186	1-85	gi—1222069	3-oxoacyl-[acyl-carrier protein] reductase [Haemophilus influenzae]
149.4	10-191	1-85	gi—309860	beta-hydroxysteroid dehydrogenase [Comamonas testosteroni]
149.1	14-196	1-85	gi—912437	7alpha-hydroxysteroid dehydrogenase [Escherichia coli]
148.0	10-192	1-85	gi—1419053	unknown [Mycobacterium tuberculosis]
145.5	325-504	1-85	gi—695398	hydratase-dehydrogenase-epimerase [Candida tropicalis]
133.1	6-193	1-85	gi—975895	17-beta-hydroxysteroid dehydrogenase [Homo sapiens]
127.7	8-235	1-85	gi—181037	carbonyl reductase [Homo sapiens]
116.6	37-222	1-85	gi—179475	11-beta-hydroxysteroid dehydrogenase [Homo sapiens]
115.4	32-213	1-85	gi—1054531	11-cis-retinol dehydrogenase [Bos taurus]
90.6	86-188	14-82	gi—304662	alcohol dehydrogenase [Drosophila immigrans]
65.8	118-244	12-85	gi—957251	oxidoreductase required for shoot apex development=FEY [A. thaliana]
23.2	138-195	46-85	gi—46868	ORF3 protein [Streptomyces coelicolor]
21.3	32-203	2-58	gi—861340	similar to ribitol dehydrogenase [Caenorhabditis elegans]
19.7	15-101	1-22	gi—603171	unknown [Escherichia coli]
19.1	12-45	58-85	gi—453866	tropinone reductase homologue [Arabidopsis thaliana]
18.8	8-18	1-11	gi—145888	ORF3 [Escherichia coli]
18.5	4-41	54-85	gi—699381	glucose 1-dehydrogenase [Mycobacterium leprae]
18.0	3-157	1-60	gi—473600	dTDP-glucose dehydratase [Streptomyces fradiae]
17.7	1-33	59-85	gi—1053075	orf1; similar to E.coli EnvM [Proteus mirabilis]
17.6	128-184	47-85	gi—641817	halohydrin epoxidase A [Corynebacterium sp.]
17.3	19-168	1-77	gi—641819	halohydrin epoxidase B [Corynebacterium sp.]
17.3	4-14	1-11	gi—415277	unknown [Escherichia coli]
16.7	1-13	73-85	gi—887852	ORF_f67p [Escherichia coli]
16.0	1-26	66-85	gi—1234827	Orf1; similar EnvM [Legionella pneumophila]
15.6	262-313	51-85	gi—237650	enoyl-acyl carrier protein reductase [Brassica napus]
15.3	85-149	27-67	gi—1332595	dNDP-glucose dehydratase [Streptomyces sp.]
14.5	28-147	2-40	gi—618456	norsolornic acid [Aspergillus parasiticus]
13.9	10-20	1-11	gi—471145	ORFUP [Sphingomonas paucimobilis]
13.7	1-13	73-85	gi—1166429	K08F4.9 [Caenorhabditis elegans]
13.4	217-298	25-85	gi—1055124	coded for by C. elegans cDNA yk62b4.3
13.0	98-173	27-67	gi—1314581	dTDP-D-glucose-4,6-dehydratase [Sphingomonas S88]
13.0	89-143	29-59	gi—1359482	dNDP-glucose dehydratase [Amycolatopsis mediterranei]
12.8	97-171	27-67	gi—398120	TDP-glucose oxireductase [Xanthomonas campestris]
12.4	9-117	1-37	gi—1143392	uridine diphosphate glucose epimerase [Arabidopsis thaliana]
12.2	113-186	50-85	gi—203979	dihydropteridine reductase (EC 1.6.99.7) [Rattus norvegicus]
12.2	116-189	50-85	gi—181553	dihydropteridine reductase (EC 1.6.99.7) [Homo sapiens]
12.0	101-174	27-67	gi—1001273	hypothetical protein [Synechocystis sp.]
10.8	2-22	68-85	gi—666992	alcohol dehydrogenase [Drosophila mojavensis]
10.4	6-164	2-67	gi—413996	ipa-72d gene product [Bacillus subtilis]
10.3	25-200	1-19	gi—506333	HrEpiB [Halocynthia roretzi]
9.9	8-116	1-37	gi—1173555	UDP-galactose-4-epimerase [Pisum sativum]
9.6	3-93	1-27	gi—567874	thymidine diphosphoglucose 4,6-dehydratase [S. erythraea]
9.6	1-15	71-85	gi—516105	aklaviketone reductase [Streptomyces sp.]
9.4	23-64	37-59	gi—699306	hypothetical protein [Mycobacterium leprae]
9.3	2-29	1-18	gi—1294775	ADP-L-glycero-D-manno-heptose-6-epimerase [Haemophilus influenzae]
8.9	3-111	1-41	gi—1429254	UDP-glucose 4-epimerase [Bacillus subtilis]
8.9	4-154	1-58	gi—406095	UDP-glucose 4-epimerase [Neisseria meningitidis]
8.9	3-62	1-15	gi—294198	malate dehydrogenase [Photobacterium sp.]
8.6	38-48	1-11	gi—466869	gpdB; B1496_F1_31 [Mycobacterium leprae]
8.5	87-198	13-85	gi—407314	inhA peptide (AA 1-269) [Mycobacterium tuberculosis]
8.5	87-198	13-85	gi—1155270	enoyl ACP reductase [Mycobacterium bovis]
8.5	58-95	33-56	gi—1381396	ribulose bisphosphate carboxylase/oxygenase large subunit

Table 2: **Selected Meta-MEME output from from an analysis of Genpept96.** We show some high scoring sequences that contain all 85 residues in the six motifs. In 3-ketoacyl-acyl carrier protein reductase, these map to residues 8-188 with a score of 178.7. In carbonyl reductase [41], these map to residues 8-235 with a ascore of 127.7. Motif residues 14-82 map to residues 86-188 on *Drosophila immigrans* alcohol dehydrogenase with a score of 90.6. Analysis of proteins with scores from 23.2 to 8.5 bits reveal that the first protein that is not a member of the short chain dehydrogenase family is malate dehydrogenase with a score of 8.9 bits, followed by ribulose bisphosphate carboyxlase/oxygenase with a score of 8.5 bits. The sequences of several homologs, such as halohydrin epoxidase [42] and the sugar epimerases [43, 44, 45], have diverged from the signature motif used in PROSITE [46], which has made identification of their ancestry difficult.

combined the first six motifs for each protein into a single sequence, which by virtue of the MEME analysis can be aligned with the other thirty-six proteins. Two equally parsimonious phylogenies were discovered by Phylip [26]. One of these two is shown in Figure 4; the other phylogeny was similar. Phylogenies using the entire sequences of 11β-hydroxysteroid dehydrogenase-type 1, 17β-hydroxysteroid dehydrogenase-types 1, 2, and 3, and β-hydroxybutyrate dehydrogenase [23], as well as bacterial steroid dehydrogenases [22] have been determined previously [23] and are in general agreement with that from the motifs. In particular, the type 1 11β- and 17β-hydroxysteroid dehydrogenases cluster together on a branch separate from 17β-hydroxysteroid dehydrogenase-type 2, which clusters with β-hydroxybutyrate dehydrogenase. On a separate branch is 17β-hydroxysteroid dehydrogenase-type 3. Thus, the information in the eighty-five residues in the first six motifs gives a useful phylogeny for the steroid dehydrogenases.

4 Discussion

There is a strong biological basis for the sensitivity of Meta-MEME. Motifs 1 and 2 are part of the nucleotide cofactor binding site [47, 48, 49]; motif 3 contains the catalytic site. A protein sequence that had motifs 1 and 3 interchanged would not have the same 3D structure and could not function the way the steroid dehydrogenases do. By scoring protein similarity and dissimilarity on the basis of motif order and spacing, the HMM method is using the spatial information in the 3D structure of the canonical dehydrogenase to identify homologs from the noise of unrelated proteins that have islands of amino acid sequence similarity to the dataset. Comparisons of protein 3D structures is the most sensitive method for determining homology [50], which we propose explains the excellent ability of HMM to separate homologs from noise as seen in Figure 3.

In summary, the combination of HMM and MEME into the Meta-MEME tool provides a sensitive and selective method for homology searches to identify distantly related proteins. This facilitates collecting large and diverse collections of homologous proteins for motif analysis for use in elucidating the relationship between structure, function and evolution.

Acknowledgments

William Grundy is funded by the National Defense Science and Engineering Graduate Fellowship Program. Charles Elkan is funded by a Hellman Faculty Fellowship from UCSD. Timothy Bailey is supported by the National Biomedical Computation Resource, an NIH/NCRR funded research resource (P41 RR-08605), and the NSF through cooperative agreement ASC-02827. Paragon time was made available through a grant to NBCR (NIH P41 RR08605).

References

[1] R. D. Fleishmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bultd, J-F. Tomb, B. A. Dougherty, and J. M. Merrick et al. *Science*, 269:496–512, 1995.

[2] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleishman, C. J. Bult, A. R. Kerlavage, G. Sutton, and J. M. Kelley et al. *Science*, 270:397–403, 1995.

[3] E. V. Koonin, A. R. Mushegian, and K. E. Rudd. *Current Biology*, 6:404–416, 1996.

[4] W. R. Pearson and D. J. Lipman. *Proceedings of the National Academy of Sciences of the United States of America*, 85:2444–2448, 1988.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[6] M. Gribskov, A. D. MacLachlan, and D. Eisenberg. *Proceedings of the National Academy of Sciences of the United States of America*, 84:4355–4358, 1987.

[7] K. Rohde and P. Bork. *Computer Applications in the Biosciences*, 9:183–189, 1993.

[8] S. Henikoff and J. G. Henikoff. *Nucleic Acids Research*, 19:6565–6572, 1991.

[9] T. L. Bailey and C. P. Elkan. The value of prior knowledge in discovering motifs with MEME. In C. Rawlings, D. Clark, R. Altman, L. C. Hunter, and L. C. Rawlings, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 21–29. AAAI Press, 1995.

[10] W. N. Grundy, T. L. Bailey, and C. P. Elkan. ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Computer Applications in the Biosciences*, 12(4):303–310, 1996.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[12] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[13] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3):1059–1063, 1994.

[14] S. L. Baldauf, J. D. Palmer, and W. F. Doolittle. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*, 93:7749–7754, 1996.

[15] V. Laudet, C. Hanni, J. Coll, F. Catzelflis, and D. Stehelin. Evolution of the nuclear receptor gene superfamily. *EMBO Journal*, 11:1003–1013, 1992.

[16] M. E. Baker. *Steroids*, 56:354–360, 1991.

[17] B. Persson, M. Krook, and H. Jornvall. *European Journal of Biochemistry*, 200:537–543, 1991.

[18] G. M. Tannin, A. K. Agarwal, C. Monder, M. I. New, and P. C. White. *Journal of Biological Chemistry*, 266:16653–16658, 1991.

[19] Z. Krozowski. *Molecular and Cellular Endocrinology*, 84:C25–C31, 1992.

[20] H. Jornvall, B. Persson, M. Krook, S. Atrian, R. Gonzalez-Duarte, J. Jeffry, and D. Ghosh. *Biochemistry*, 34:6003–6013, 1995.

[21] M. E. Baker. *Biochemistry Journal*, 300:605–607, 1994.

[22] M. E. Baker. Sequence analysis of steroid and prostaglandin metabolizing enzymes: application to understanding catalysis. *Steroids*, 59:248–258, 1994.

[23] M. E. Baker. Unusual evolution of mammalian 11β- and 17β-hydroxysteroid and retinol dehydrogenases. *Bioessays*, 18:63–70, 1996.

[24] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1995.

[25] S. R. Eddy. Multiple alignment using hidden Markov models. In C. Rawlings et al., editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120. AAAI Press, 1995.

[26] J. Felsenstein. PHYLIP — phylogeny inference pacakge (version 3.2). *Cladistics*, 5:164–166, 1989.

[27] D. Ghosh, Z. Wawrzak, C. M. Weeks, W. L. Duax, and M. Erman. *Structure*, 2:629–640, 1994.

[28] K. I. Varughese, N. H. Xuong, P. M. Kiefer, D. A. Matthews, and J. M. Whiteley. *Proceedings of the National Academy of Sciences of the United States of America*, 91:5582–5586, 1994.

[29] R. Breton, D. Housset, C. Mazza, and J. C. Fontecilla-Camps. *Structure*, 4:905–915, 1996.

[30] J. B. Rafferty, J. W. Simon, C. Baldock, P. J. Artymiuk, P. J. Baker, A. R. Stuitje, A. R. Slabas, and D. W. Rice. *Structure*, 3:927–938, 1995.

[31] A. Banerjee, E. Dubnau, A. Quemard, V. Balasumbramanian, K. S. Um, T. Wilson, D. Collins, G. de Lisle, and W. R. Jacobs. *Science*, 263:227–230, 1994.

[32] N. Tanaka, T. Nonaka, T. Tanabe, T. Yoshimoto, D. Tsuru, and Y. Mitsui. *Biochemistry*, 35:7715–7730, 1996.

[33] J. Obeid and P. C. White. *Biochemistry and Biophysics Research Communications*, 188:222–227, 1992.

[34] R. Albalat, R. Gonzalez-Duarte, and S. Atrian. *FEBS Letters*, 308:235–239, 1992.

[35] Z. Chen, J. C. Jiang, Z. G. Lin, W. R. Lee, M. E. Baker, and S. H. Chang. *Biochemistry*, 32:3342–3346, 1992.

[36] T. J. Puranen, H. Poutanen, H. E. Peltoketo, P. T. Vihko, and R. K. R. K. Vihko. *Biochemistry Journal*, 304:289–293, 1994.

[37] H. M. Wilks and M. P. Timko. *Proceedings of the National Academy of Sciences of the United States of America*, 92:724–728, 1995.

[38] S. W. Chenevert, N. G. Fossett, S.H. S. H. Chang, I. Tsigelny, M.E. M. E. Baker, and W. R W. R. Lee. *Biochemistry Journal*, 308:419–423, 1995.

[39] I. Tsigelny and M. E. Baker. *Biochemistry and Biophysics Research Communications*, 21:859–868, 1995.

[40] I. Tsigelny and M. E. Baker. *Journal of Steroid Biochemistry and Molecular Biology*, 55:589–600, 1995.

[41] B. Wermuth. *Prostaglandins*, 44:5–9, 1992.

[42] F. Yu, T. Nakamura, W. Mizunashi, and I. Watanabe. *Bioscience, Biotechnology, Biochemistry*, 58:1451–1457, 1994.

[43] L. Holm, C. Sander, and A. Murzin. *Nature Structural Biology*, 1:146–147, 1994.

[44] G. Labesse, A. Vidal-Cros, J. Chomilier, M. Gaudry, and J. P. Mornon. 1994.

[45] M. E. Baker and R. Blasco. *FEBS Letters*, 301:89–93, 1992.

[46] A. Bairoch. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research*, 20:2013–2018, 1992.

[47] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland, 1991.

[48] R. K. Wierenga, M. C. De Maeyer, and W. G. J. Hol. Interaction of pyrophosphate moieties with α-helices in dinucleotide binding proteins. *Biochemistry*, 24:1346–1357, 1985.

[49] R. K. Wierenga, P. P. Terpstra, and W. G. J. Hol. Prediction of the occurrence of the ADP-binding β-α-β-fold in proteins using an amino acid sequence fingerprint. *Journal of Molecular Biology*, 187:101–107, 1986.

[50] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5:823–826, 1986.