

An investigation of distant homology detection methods for multidomain protein families

Andrey Rzhetsky

Columbia Genome Center and
Department of Medical Informatics
Columbia University
ar345@columbia.edu

William Noble Grundy

Department of Computer Science
Columbia University
bgrundy@cs.columbia.edu

Reina E. Riemann

Columbia Genome Center
Columbia University
rer20@columbia.edu

Andrea Califano

IBM Computational Biology Center
TJ Watson Research Center
acal@us.ibm.com

Abstract

Homology detection methods first emerged as pairwise alignment algorithms, such as Needleman-Wunsch, Smith-Waterman, FASTA, and BLAST. Although these pairwise comparison methods continue to play the major role in heavy-duty genomics applications, there are a few recent more sensitive algorithms that rely on statistical models built from training sets of related sequences.

Many interesting methods fit in this new category. The majority are based on hidden Markov models (HMMs), on position-specific scoring matrices (PSSMs), or on clustering approaches that extend the original pairwise analysis algorithms. To further complicate matters, there are a variety of methods that are capable of discovering and building protein domain models by analyzing families of related sequences.

Given the rapid proliferation of these techniques, it is important to establish a sound methodology to uniformly compare the performance of such algorithms in a reasonable and objective fashion.

The objective of this paper is twofold. It introduces a relatively simple framework for the comparative analysis of consensus-based homology detection algorithms and for the domain model discovery algorithms. It also introduces a novel PSSM-based technique called SPLASHSearch that compares favorably with other established techniques.

To test these algorithms, training sets have been built from sequences that include related protein domains. Only subsequences that include the domain are used for domain model building and training, in order to avoid contamination of the tests from multiple, diverse domain signatures. The goal is to determine the ability of the various techniques to perform rapid, accurate sequence annotation given a correct domain definition by a set of training sequences.

Three key observations arise from this exercise. First, algorithms based on consensus motifs are becoming as accurate and sensitive as the more computationally expensive methods based on HMMs. Second, the ability to identify motifs impacts the performance of the search methods. Finally, for the HMM method prepro-

cessing the sequences with a multiple sequence alignment algorithm has a negative impact on the search performance for some “difficult” domains.

Keywords: Homology detection, proteins, motif analysis

Introduction¹

Within the young field of computational biology, the protein homology detection task is a classic problem (Barsalou & Brutlag 1991). The task consists of finding, for a given protein or protein family, all distantly related protein sequences in a large database of unannotated sequences. Two sequences that share a common evolutionary ancestor are said to be homologous. Because we do not have access to ancestral protein sequences, the homology detection task is necessarily inferential.

Initially, inferences of homology were made based upon pairwise comparisons of protein sequences, using dynamic programming algorithms such as the Needleman-Wunsch (Needleman & Wunsch 1970) and Smith-Waterman (Smith & Waterman 1981) algorithms. The popular database search tools BLAST (Altschul *et al.* 1990) and FASTA (Pearson 1985) are fast approximations of these dynamic programming algorithms.

Unfortunately, many “difficult” protein sequences exhibit no significant pairwise similarity to previously characterized sequences in public databases. However, evidence of significant homology of the “difficult” sequences to known sequences often emerges from comparison of the “difficult” sequences to statistical models describing families of known sequences. In this context, the ability to build complex consensus models from a training set of proteins that share a common function or structure is a powerful tool. There

¹Corresponding author: Andrey Rzhetsky, Department of Medical Informatics, Russ Berrie Pavilion, Unit 109, 1150 St. Nicholas Ave., Tel: (212) 304-7552, Fax: (212) 304-5515

are three key reasons for this power. First, in a consensus model, protein regions that are less functionally constrained contribute far less to the sequence-to-model comparison than in a pairwise sequence analysis. This tends to reduce the impact of “noisy” sequences that may contain multiple diverse functional signatures. Second, rather than using a generic model of amino acid substitution (Schwartz & Dayhoff 1978; Henikoff & Henikoff 1992), more accurate family- and site-specific models can be used. Finally, pairwise sequence analysis often assumes that sequences change mostly through amino acid substitution and insertion/deletion. However, in reality, proteins domains often can be swapped without a loss of protein function, and a domain exchange among non-homologous proteins is an important instrument of molecular evolution.

The PROSITE database (Bairoch 1991) represents the first systematic attempt to identify biologically relevant consensus sequences, or motifs, in functionally related families. Once identified, motifs can be easily represented as PSSMs. Over the past five years, a number of statistical and deterministic motif discovery algorithms have been introduced in the literature. These include MEME (Bailey & Elkan 1994), SPLASH (Califano 2000), the Gibbs sampler (Neuwald, Liu, & Lawrence 1995), Teiresias (Rigoutsos & Floratos 1998), PRATT (Jonassen, Collins, & Higgins 1995), and E-MOTIF (Nevill-Manning, Wu, & Brutlag 1998).

The goal of this paper is to propose an objective framework for the comparison of consensus-based homology detection algorithms and of the underlying motif discovery algorithms. This framework allows for the analysis of the accuracy and computational efficiency of these approaches. The comparison methodology, training sequences, and results are available on the web, thereby simplifying the inclusion of new algorithms and new data sets over time.

Clearly, given the large number of available algorithms, an exhaustive analysis of all available methods is impossible. Therefore, we have initially selected four independent algorithms. Three of them — HMMER, MAST, and Meta-MEME — are publicly available. The fourth, SPLASHSearch, is an additional contribution of this paper and compares favorably with the other methods. HMMER has been tested both with training sets formed by unaligned sequences and with ones that had been previously globally aligned using CLUSTALW (Thompson, Higgins, & Gibson 1994a).

With the exception of HMMER, all methods rely on a set of PSSM models for the given functional sequence training set. In combination with these methods we have tested two independent motif discovery algorithms. The first is MEME, which is based on a probabilistic, maximum likelihood model. The second is SPLASH, which is based on a deterministic pattern discovery algorithm. SPLASH has been recently shown to offer an extremely efficient, yet exhaustive way to discover conserved motifs based on amino acid similar-

ity metrics. It has also been shown to identify, with high probability, protein regions that are biologically relevant and highly conserved. In an exhaustive test against the PROSITE motif database (Hart *et al.* 2000) SPLASH produced motifs that were highly overlapping with PROSITE reported motifs in more than 75% of the cases. MEME, on the other hand, is the standard algorithm used to generate PSSM models for two of the four tested methods. Other choices are possible and will be added to this benchmark in successive iterations.

Finally, Meta-MEME has been tested both with trained and untrained (uniform) transition probabilities between the individual PSSM models.

This experimental setup yields a total of nine independent techniques: HMMER (aligned/unaligned), MAST with either MEME- or SPLASH-based PSSMs, Meta-MEME (trained/untrained) with either MEME- or SPLASH-based PSSMs, and SPLASHSearch with SPLASH-based PSSMs. The latter method extends the p -value integration method pioneered in MAST with a probabilistic model for the relative position of PSSM matches on the sequences. Since this additional probability model is not generated by MEME, it has not been possible to study the performance of SPLASHSearch with MEME-based PSSMs.

Training sets in our analysis comprise individual protein domains rather than complete multi-domain proteins. This choice allows for a clear definition of a set of database sequences bearing genuine homology to the training sequences. The alternative approach, involving discovering and training domain models from complete multidomain protein sequences, would often result in statistical models reflecting more than one protein domain. Since pairs of multi-domain proteins frequently exhibit only partial homology, the training set is likely to contain domains that are not shared by all sequences in the set. Furthermore, the “true positive” database matches to such training set would have to include all sequences having at least one of the domains represented in the training set. Note that the majority of currently existing methods are not designed to deal with domain models trained from multi-domain data sets.

The domain sequences have been obtained by an automated methods that analyzes the annotations of sequences in SWISSPROT 38. Five types of domains have been chosen for this exercise, including two simple domains with similarities identifiable with pairwise comparison approaches and three complex domains with lower global conservation that is often undetectable with pairwise comparison methods.

The protein domain selection procedure is fairly laborious and requires some manual supervision to avoid potential contamination of the training sets and of the lists of true positive and true negative examples with spurious data. Therefore, the initial benchmark is limited to five domains. This should be considered an initial set that will grow over time to include more complex and discriminative examples. Clearly, this reduced set does not allow us to draw definitive conclusions about the

relative performance of the various techniques. However, the results show a significant range of variability and several definite trends. With the exception of one or two of the simpler domain families that were easily identified by all methods, more complex domains yield a wide range of performance, with differences as large as 50%.

In general, three observations emerge. First, the motif-based approaches offer accuracy and sensitivity that is comparable to that of the more computationally expensive HMM-based algorithms at a fraction of the computational cost. In particular, SPLASHSearch and HMMER unaligned offer the best results, with SPLASHSearch winning by about 15% in one of the 5 categories and tying in the remaining four. HMMER, however, is about an order of magnitude more computationally expensive. The MAST+SPLASH combination follows closely with a loss by 0.5% and one by 6.5% respectively. This approach is about 1.5 times faster than SPLASHSearch. MAST+MEME follows with four losses by 2%, 30%, 10%, and 11%. Finally, results for Meta-MEME in combination with MEME and SPLASH are mixed but, in general, it does not perform as well as the other techniques. It is likely that this algorithm would perform better with training sets composed of complete sequences and with additional fine-tuning of default parameters.

Second, different choice of motif discovery methods can result in significant difference in terms of accuracy and sensitivity. In particular, MAST with SPLASH-discovered motifs outperformed the same algorithm with MEME-discovered motifs in 4 out of 5 categories.

Finally, training HMMER domain models with CLUSTALW-aligned sequences produces inferior results when the training set sequences belong to complex families that have low sequence conservation. For the most complex dataset, the immunoglobulin-like domains, performance drops by nearly 50% when CLUSTALW is used in conjunction with HMMER. This confirms that the substitution/deletion/insertion model of evolution, which is enforced in multiple sequence alignment, is inappropriate when complex functional signatures are involved.

Methods

Analyses were performed using five protein families extracted from the SWISSPROT database, version 38 (Bairoch 1994). Each family is defined by the presence of a particular type of domain or motif: CUB, c1q, galaptin, immunoglobulin and kringle. CUB domains are found in two types of proteins: proteins that regulate the developmental process and proteins in the complement system. The name ‘‘CUB’’ indicates that this domain occurs in three complement components: C1r/C1s, Uegf, and bone morphogenic protein-1. The c1q domain participates in complement formation in vertebrates and occurs in the c1q subunit of the C1 enzyme complex. Galaptin domains are common to a class of soluble, developmentally regulated proteins called

Family	Seqs	Sequence	Domains	Domain
		AAAs		AAAs
CUB	36	22 878	63	7129
c1q	27	10 706	13	1677
galaptin	40	8048	15	2122
IG	386	258 467	500	37 469
kringle	40	27 617	52	4178

Table 1: **Protein families investigated.** The columns contain the names of the families, number of sequences in each family, number of amino acids in each family, number of domains used in training, and the number of amino acids in those domains.

galactoside-binding lectins. These proteins are synthesized by all vertebrates and seem to be involved in differentiation, cellular regulation and tissue construction. The immunoglobulin domain is common among a large superfamily of proteins, the most famous representatives of which are the antibodies in the human immune system. Immunoglobulins participate in specific recognition of other large molecules. Finally, kringles are triple-looped, disulfide, cross-linked domains found in a varying numbers of copies in some serine proteases and plasma proteins. Kringle domains are thought to play a role in binding mediators, such as membranes, other proteins or phospholipids, and in the regulation of proteolytic activity. More information about these families is available at <http://www.expasy.ch>. Of the five families investigated, the first two are relatively closely clustered in sequence space, whereas the last three families contain highly dissimilar sequences whose global homology is nearly undetectable by pairwise sequence comparison methods (data not shown).

Lists of sequences in each family were derived from SWISSPROT annotations by searching for corresponding keywords. Single domains were extracted from the sequences by exploiting the ‘‘motif’’ and ‘‘domain’’ features in SWISSPROT. For each family, the training set consists of approximately half the domains in the family. Some characteristics of these families are described in Table 1. The family members and training sets are available at <http://www.cs.columbia.edu/~bgrundy/5-families>.

For each family, MEME and SPLASH each discover one set of motifs. MEME, version 2.2.2, employs the default parameter settings from the web interface (<http://www.sdsc.edu/MEME>). These settings include Dirichlet mixture priors (Brown *et al.* 1995) weighted by the megaprior heuristic (Bailey & Gribskov 1996), a minimum motif width of 12 and a maximum of 55. The motif model is a fully general, two-component mixture model that allows for a given motif to appear any number of times in the given sequences. MEME discovers a total of ten motifs for each family. SPLASH, version 1.0, also uses its default parameter settings to discover ten conserved motifs, as described in (Hart *et al.* 2000). Motifs are progressively discovered and masked

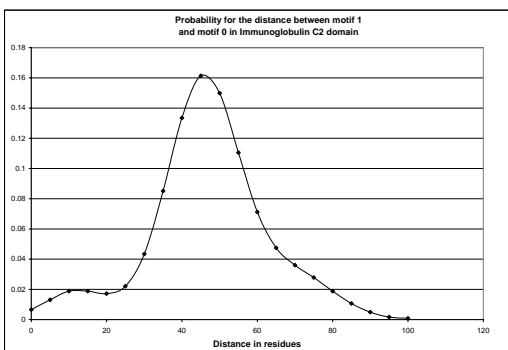


Figure 1: **Relative motif distance probability for the first and second most likely motifs in the Immunoglobulin family.**

in the sequences, starting with the most statistically significant ones, so that the motifs are guaranteed to be non-overlapping. In particular, a BLOSUM50 matrix (Henikoff & Henikoff 1992) with a threshold of 0 is used as the similarity metric, one identity match is required, and reported motifs must have at least four tokens. The density constraint requires three matching tokens in any window of length 12. PSSMs are built using the amino acid counts for all the locations where the reported motifs occur in the training set. Pseudo-counts are computed using the minimal risk approach (Wu, Nevill-Manning, & Brutlag 1999).

The motifs are used in three different ways to search the SWISSPROT database, version 38. The first database search program, MAST (Bailey & Gribskov 1998), is part of the MEME software distribution package. MAST scans each database sequence with each motif separately and then combines the highest-scoring matches. MAST computes E-values by assuming that the motif matches within a given database sequence are independent of one another. MAST can search using motifs discovered either by MEME or SPLASH.

The second program, SPLASHSearch, extends the MAST multiple p -value approach by taking into account the relative spacing of motifs, based on a probabilistic model. The model is computed empirically from the training set by determining the probability $p(m_j|m_i, d)$ of observing a motif m_j at distance d residues from motif m_i . This function is smoothed using a gaussian filter with $\sigma = 5$ and renormalized. Figure 1, for instance, plots the final value of $p(m_0|m_1, d)$ for the two PSSM models with the highest incidence in the Immunoglobulin C2 domain family. These correspond to the two SPLASH regular expressions [ILMFV] G.Y.C and G. [ILMFV] . [ILMFV] . C, which have been underlined, as an example, in two sequence fragments:

P04217: . . . SLLKPLANVTLTCQARLETPD FQLFKNGVAQEPVHLDSPA IKHQFLLTGDTQGRYRC RSGL . . .

P15364: GQNLLELTCHANGFPKPTISWAREHNAVMPAGHLLAEPTLRIRSVHRMDRGGYYCIA QNGEGQPD

This information is then used in the context of an evidence integration framework such that motifs that occur in unlikely configurations are weighted independently rather than jointly, while motifs that appear in a configuration similar to one observed in the training set contribute jointly to the final score with the weighed product of their p -values.

This weighting is accomplished as follows. Let us assume we have a set of motifs matches $\{m_i\}$, where motif m_i occurs at position l_i on a given sequence and has a p -values p_i . SPLASHSearch computes

$$\max_i \left[-\log(p_i) + \sum_{j \neq i} [-\log(p_j) * p(m_j|m_i, l_j - l_i)] \right].$$

Therefore, if $p(m_j|m_i, l_j - l_i)$ is close to zero for all js , the score is dominated by the term $\max[-\log(p_i)]$. On the other hand, if there is an i such that $p(m_j|m_i, l_j - l_i) \gg 0$ for some js , then the contribution from the second term in the sum becomes not negligible. SPLASHSearch uses motifs discovered by SPLASH, but not by MEME because these additional probability density models are not produced by the latter algorithm. In the future, however, the output from MEME could be used to determine the probabilistic model in an intermediate step.

The third search method uses Meta-MEME, version 2.1, to build motif-based HMMs. Each set of motifs is assembled into a motif-based HMM. The model topology is completely connected, meaning that transitions are included in the model from the end of every motif to the beginning of every other motif. This topology allows motifs to appear in any order and to appear multiple times within a single sequence. Sequence regions between motifs are modeled using a single “free insertion module” (Hughey & Krogh 1996); i.e., a state with one self-transition and one outgoing transition, both with a transition “probability” of 1.0. Initially, all motif-to-motif transition probabilities are initialized uniformly. Such a model is referred to below as an “untrained” model. The motif-to-motif transition probabilities can be trained using expectation-maximization. Both the untrained and trained models are used to search the protein database. The Meta-MEME search tool, `mhmms`, computes the probability of the most likely path through the model. This score is reported as a log-odds scores in bits, using a simple, zero-order Markov chain as the background model (Grundy 1998). Meta-MEME can build models from motifs discovered by MEME or SPLASH.

For comparison with the motif-based methods, homologs are also detected using profile HMMs. Two sets of HMMs are constructed by HMMER (Eddy 1995). The first set is derived by HMMER, version 2.0, from multiple alignments of each family, as produced by CLUSTALW (Thompson, Higgins, & Gibson 1994b) using neighbor-joining trees (Saitou & Nei 1987). The

Method	Time
Motif discovery by SPLASH	< 1 minute
Motif discovery by MEME	30 minutes
Model building and search by HMMER	20 minutes
Database search by MAST	< 1 minute
Database search by Meta-MEME	30 minutes
Database search by SPLASHSearch	< 1 minute

Table 4: **Comparison of computational requirements of motif discovery and homology detection methods.** Times are averages for the 5 families. Notable exceptions are reported in the text.

second set of profile HMMs is built directly from the unaligned domain sequences using HMMER, version 1.8. Model calibration and database searching are carried out for both types of models using HMMER 2.0. For each program, the default parameter settings are used.

Each homology detection method produces as output a ranked list of putative homologs. These lists are summarized using receiver operating characteristic (ROC) analysis (Gribskov & Robinson 1996). The ROC score is the area under a curve that plots true positives versus false positives for varying score thresholds. ROC analysis combines measures of a search’s sensitivity and selectivity. The ROC_{100} score is the area under the ROC curve, up to the first 100 false positives. ROC_{100} scores are normalized to range from 0 to 1, with 1 corresponding to the most sensitive and selective search.

In total, nine homology detection methods were investigated, as summarized in Table 2. None of the parameter settings was varied prior to reporting these results. Hence, it is likely that alternate values would provide better homology detection performance. In practice, however, the correct set of homologs is not known *a priori*, and parameter optimization is therefore not possible.

Results and discussion

The results of our experiments are summarized in Table 3. Between the two motif discovery algorithms, MEME and SPLASH, neither consistently provides superior homology detection performance, but SPLASH appears to be on average significantly faster than MEME (the exception is analysis for the set of full-length CUB-containing protein sequences). Also, in combination with MAST, it outperforms MEME in 4 out of the 5 tests. Among the five database search techniques, MAST, SPLASHSearch and HMMER all provide excellent performance, although the MAST and SPLASHSearch search programs are significantly faster than their HMMER and Meta-MEME counterparts.

Motif discovery and search techniques

In combination with MAST, SPLASH is a winner over MEME in 4 out of 5 categories, in two of these by more than 5%. When these algorithms are used in combina-

Method	CUB	clq	gal	IG	krin
MAST	—	s	S	s	S
Meta-MEME (un)	M	S	M	s	M
Meta-MEME (tr)	S	—	M	S	S

Table 5: **No motif discovery algorithm provides significantly superior homology detection performance.** The table lists, for each of three database search methods, the motif detection program that provides better relative performance (S=SPLASH, M=MEME). A lowercase letter is used when ROC_{100} values are better by more than 0.01 but less than 0.05. When two searches produced ROC_{100} values within 0.01 of one another, no winner is declared.

tion with Meta-MEME, results are more mixed and no clear trend emerges. In general, however, the combination of MAST and SPLASH seems to offer the best overall accuracy coupled with the best overall computational speed.

Table 5 compares the homology detection performance provided by motifs discovered by these two programs. Analysis of the domain models produced by SPLASH and MEME indicate that SPLASH tends to generate motif models that more completely cover the training sequences than do the models produced by MEME.

Surprisingly, estimation of non-uniform transition probabilities between PSSMs by either SPLASH or MEME does not provide any apparent advantage in selectivity of search (Tables 3, 5), even though this technique is more expensive computationally (data not shown). A plausible explanation for this result is that the training set has to be significantly larger to provide for reliable estimates of these parameters, and the currently obtained estimates include large sampling errors. It appears that with the current size of training datasets uniform transition probabilities appear to be the better choice in PSSM-based searches.

Furthermore, comparison of the overall performance of search methods (Tables 3) with their computation cost (Tables 4) indicates that MAST and SPLASHSearch provide a near-optimum tradeoff between database search quality (sensitivity and selectivity) and search computational complexity.

In Table 6 we report the results of motif discovery, using SPLASH and MEME, for the C2 domain of the Immunoglobulin family. The most notable difference between the two motif discovery methods is that motifs reported by MEME seem to be shorter and have higher rates of occurrence than those reported by SPLASH. The occurrence rates reported by SPLASH are somewhat misleading because SPLASH reports regular expression matches rather than the expected number of occurrences of the corresponding PSSM model. Once generated from the corresponding regular expression, SPLASH motifs usually have a significantly higher incidence in the dataset. For instance, the first regular

	Model building	Search tool	Scores
1	MEME	MAST	E-value
2	MEME + Meta-MEME (untrained)	mhms	Viterbi log-odds
3	MEME + Meta-MEME (trained)	mhms	Viterbi log-odds
4	SPLASH	MAST	E-value
5	SPLASH + Meta-MEME (untrained)	mhms	Viterbi log-odds
6	SPLASH + Meta-MEME (trained)	mhms	Viterbi log-odds
7	SPLASH	SPLASHSearch	see text
8	HMMER from unaligned sequences	hmmsearch	Viterbi log-odds
9	HMMER from CLUSTALW alignment	hmmsearch	Viterbi log-odds

Table 2: **Summary of homology detection methods.** See the text for a detailed description of each method.

Method	CUB	c1q	galaptin	IG	kringle
MEME + MAST	1.0000	0.8371	0.6013	0.7114	0.8937
MEME + Meta-MEME (untrained)	0.9783	0.7392	0.5844	0.6362	0.9793
MEME + Meta-MEME (trained)	0.1666	0.8571	0.9279	0.0455	0.2468
SPLASH + MAST	1.0000	0.8571	0.9938	0.7434	1.0000
SPLASH + Meta-MEME (untrained)	0.0116	0.8407	0.3762	0.6574	0.6581
SPLASH + Meta-MEME (trained)	0.2349	0.8557	0.7775	0.9015	0.9100
SPLASH + SPLASHSearch	1.0000	0.8571	0.9989	0.8119	1.0000
HMMER from unaligned sequences	1.0000	0.8571	0.8620	0.8178	1.0000
HMMER from CLUSTALW alignment	1.0000	0.8571	0.8275	0.3033	1.0000

Table 3: **Comparison of homology detection methods.** Values reported are ROC₁₀₀ scores.

SPLASH				
Id	Seq. Support	ZScore	Width	Motif
1	288	1.202e+62	14	[ILMFV].....G.Y.C
2	221	3.030e+30	8	G...[ILMFV].[ILMFV].C
3	102	6.595e+25	9	G.P.P.[ILMFV].[FWY]
4	12	1.374e+30	36	[DQEK].....L.Q....[ANST]...S.[ILV][NST]...[DEK][NDS][RQK][REK]...[DQK][RQK]
5	13	1.920e+24	23	[RQEHK][QEK]...[ILV].[LWY]T..Q[AS]...[LY]...[NT][LF]T
6	12	8.453e+17	28	[ILMV].[HPWY][ILMFV]..NG.[RQK]...[ND].....[IV][LW].....[NS]
7	12	3.080e+14	14	[DQEK]...[RQEHK][QEK][QEK]V..L[ILMV][MV][LMP]
8	12	7.579e+13	26	L[RQEHK]...F...[REHK][IL].G.[RHK].....[AS].[IFV]
9	13	1.178e+11	16	[LFWY]..DG..[ILMV]..[NDHS][NDHS][REK]...[ILMF]
10	12	1.554e+10	14	[AST].P.[ANST]..[ILMFV][QK][LFWY]S...[ANS]
MEME				
Id	Seq. Support	ZScore	Width	Motif
1	418		7	D[AS]GYTC
2	380		8	QETVTL.C
3	193		9	G.P.P.[IV]TW
4	72		10	[IV].WYK[ND]GK.L
5	7		11	LTCEV[LWM]GPTSP
6	6		6	PPAQYSWLING
7	7		11	PPA.Y.WYING
8	6		11	PPANFT[WI]QKNG
9	3		9	LSYRWLLNE
10	3		11	RQAKAVWVLP

Table 6: **Motifs Identified by SPLASH and MEME in Immunoglobulin C2 domain.**

Method	ROC ₁₀₀
MEME + MAST	0.3186
MEME + Meta-MEME (ut)	0.2584
MEME + Meta-MEME (tr)	0.3621
SPLASH + Meta-MEME (ut)	0.0000
SPLASH + Meta-MEME (tr)	0.1113

Table 7: **Comparison of domain discovery and database search methods trained from complete sequences of proteins containing at least one CUB domain.** Values reported are ROC₁₀₀ scores. The table does not contain results for HMMER package because it was not intended for discovering motifs in full-length sequences.

expression reported by SPLASH occurs 228 times in the training set. However, it largely overlaps with the motif found by MEME and, once converted into a PSSM, has a comparable incidence. The same can be said for motifs 2 and 3. These motifs also correspond to equivalent regions on the sequence to the corresponding motif identified by MEME. However, from 4 to 10, the motifs found by SPLASH are significantly longer and have better support than comparable MEME motifs. For instance MEME motifs 6, 7, and 8 are virtually identical, and the last two motifs have an incidence lower than 4. This may in part explain the better performance on the C2 domain of the MAST/SPLASH combination versus the MAST/MEME combination. In other words, while the programs are somewhat equivalent at finding highly conserved regions in the family, SPLASH has an advantage when it comes to discover motifs conserved in a small subset of the sequences. Also, overall SPLASH motifs are longer on average and may therefore have a slightly better discriminative power.

Modeling whole sequences

Discovering protein domains in full-length sequences is a difficult problem that apparently still remains open. To illustrate the difficulties associated with training domain models from full-length proteins, we present ROC₁₀₀ scores obtained from sequences containing CUB domain (Tables 7). To explain the comparison results, let us examine possible formulations of the problem of motif discovery from full-length sequences. One can formulate the goal of such analysis as identification of *all protein domains* encountered at least once in the training set. In this case, the set of true positive database matches would include all sequences having at least one of the domains encountered in the training set. This formulation would inevitably lead to attempts to build a model of a protein domain that is represented in the training set by a single sequence and therefore appears flawed. Alternatively, one can formulate the problem as identifying domains that are common for *all sequences* in the training set, which is the case with the CUB domains in our dataset, and the set of true positive database matches would include

only all CUB-containing proteins.

We have chosen the latter formulation of the problem. However, both motif discovery methods, MEME and SPLASH, apparently identify a number of motifs situated *outside* of the CUB domain. It is not surprising then that database searches with the resulting models identify a large number of “false positive” matches corresponding to training set domains other than CUB. These are indeed false positive with respect to the CUB family but they are correct matches to other domain signatures also present in members of the CUB family. This tendency to pick up large numbers of false positive database matches appears to be stronger for the SPLASH algorithm than MEME because SPLASH on average tends to give fuller coverage of the training sequences with the discovered motifs.

In conclusion we note that, currently, the largest obstacles to a rigorous comparison of domain discovery methods are the incompleteness of sequence annotations in the databases and the presence of circular logic in defining a set of true-positive database matches based on sequence annotations derived from earlier database searches.

References

- Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. 1990. A basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.
- Bailey, T. L., and Elkan, C. P. 1994. Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. In Altman, R.; Brutlag, D.; Karp, P.; Lathrop, R.; and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36. AAAI Press.
- Bailey, T. L., and Gribskov, M. 1996. The megaprior heuristic for discovering protein sequence patterns. In States, D. J.; Agarwal, P.; Gaasterland, T.; Hunter, L.; and Smith, R., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 15–24. AAAI Press.
- Bailey, T. L., and Gribskov, M. 1998. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14(1):48–54.
- Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Research* 19:2241–2245.
- Bairoch, A. 1994. The SWISS-PROT protein sequence data bank: Current status. *Nucleic Acids Research* 22(17):3578–3580.
- Barsalou, T., and Brutlag, D. 1991. Searching gene and protein sequence databases. *M. D. Computing* 8(3):144–149.
- Brown, M.; Hughey, R.; Krogh, A.; Mian, I.; Sjolander, K.; and Haussler, D. 1995. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Rawlings, C., ed., *Proceedings of the*

Third International Conference on Intelligent Systems for Molecular Biology, 47–55. AAAI Press.

Califano, A. 2000. SPLASH: Structural pattern localization analysis by sequential histograms. *Bioinformatics*.

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

Gribskov, M., and Robinson, N. L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry* 20(1):25–33.

Grundy, W. N. 1998. *A Bayesian Approach to Motif-based Protein Modeling*. Ph.D. Dissertation, University of California, San Diego, La Jolla, CA.

Hart, R.; Royyuru, A.; Stolovitsky, G.; and Califano, A. 2000. Systematic and automated discovery of patterns in PROSITE families. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*.

Henikoff, S., and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* 89:10915–10919.

Hughey, R., and Krogh, A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Computer Applications in the Biosciences* 12(2):95–107.

Jonassen, I.; Collins, J. F.; and Higgins, D. G. 1995. Finding flexible patterns in unaligned protein sequences. *Protein Science* 4:1587–1595.

Needleman, S., and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* 48:443–453.

Neuwald, A. F.; Liu, J. S.; and Lawrence, C. E. 1995. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science* 4:1618–1632.

Nevill-Manning, C. G.; Wu, T. D.; and Brutlag, D. L. 1998. Highly specific protein sequence motifs for genome analysis. *Proceedings of the National Academy of Sciences of the United States of America* 95(11):5865–5871.

Pearson, W. R. 1985. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods in Enzymology* 183:63–98.

Rigoutsos, I., and Floratos, A. 1998. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14(1):56–67.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

Schwartz, R. M., and Dayhoff, M. O. 1978. *Atlas of Protein Sequence and Structure*. Silver Spring, MD: National Biomedical Research Foundation. chapter Matrices for detecting distant relationships, 353–358.

Smith, T., and Waterman, M. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

Thompson, J. D.; Higgins, D. G.; and Gibson, T. J. 1994a. CLUSTAL W: Improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22):4673–4680.

Thompson, J. D.; Higgins, D. G.; and Gibson, T. J. 1994b. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673–4680.

Wu, T. D.; Nevill-Manning, C. G.; and Brutlag, D. L. 1999. Minimum-risk profiles of protein families based on statistical decision theory. *Journal of Computational Biology* 6(2):219–235.