Automated Identification of Putative Methyltransferases from Genomic Open Reading Frames*

Jonathan E. Katz‡, Mensur Dlakić§¶, and Steven Clarke‡‖

We have analyzed existing methodologies and created novel methodologies for the automatic assignment of *S*-adenosylmethionine (AdoMet)-dependent methyltransferase functionality to genomic open reading frames based on predicted protein sequences. A large class of the AdoMet-dependent methyltransferases shares a common binding motif for the AdoMet cofactor in the form of a seven-strand twisted β-sheet; this structural similarity is mirrored in a degenerate sequence similarity that we refer to as methyltransferase signature motifs. These motifs are the basis of our assignments. We find that simple pattern matching based on the motif sequence is of limited utility and that a new method of "sensitized matrices for scoring methyltransferases" (SM^2) produced with modified versions of the MEME and MAST tools gives greatly improved results for the *Saccharomyces cerevisiae* yeast genome. From our analysis, we conclude that this class of methyltransferases makes up ~0.6–1.6% of the genes in the yeast, human, mouse, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Escherichia coli* genomes. We provide lists of unidentified genes that we consider to have a high probability of being methyltransferases for future biochemical analyses. ***Molecular & Cellular Proteomics 2:525–540, 2003.***

S-Adenosylmethionine (AdoMet)[1]-dependent methylation is a broad class of biological processes in which a methyl group is transferred from AdoMet yielding *S*-adenosylhomocysteine and a methylated target molecule. Many of these reactions are so basic to the proper functioning of life that the lack of the gene product that performs these reactions is sufficient to severely, if not completely, abrogate the normal functioning of an organism (1–3).

Considering the broad range of target molecules, biochemists that study methylation have been blessed in that many of the AdoMet-dependent methyltransferases share common three-dimensional signatures (notably in the AdoMet binding regions) that are imperfectly reflected in similarities in their primary sequences (4). There are, at present, at least three structurally defined types of AdoMet-dependent methyltransferases. The major class (Class I) is based on a seven-strand twisted β-sheet structure (4, 5). A second recently described class (Class II) is exemplified by the SET proteins (6). Finally, a last class (Class III) is the set of membrane-associated enzymes with multiple membrane-spanning regions (7).

Herein is described the unification of developed methods to mine the information available in gene primary sequences and the screening of entire genomes in the attempt to completely assign *in silico* all known and novel AdoMet-dependent methyltransferases of the major seven-strand twisted β-sheet family. The common motifs for Class I AdoMet-dependent methyltransferases were first recognized in 1989 when three regions of similarity were noticed between the protein L-isoaspartyl *O*-methyltransferase and certain nucleic acid and small molecule methyltransferases (8). Over the years, these regions were expanded, largely by manual inspection of sequences, into Motif I, Post I, Motif II, and Motif III (9). These motifs were ultimately used for the first time in 1999 to scan the entire genome of *Saccharomyces cerevisiae* for putative methyltransferases (10). The result of the 1999 analysis was a list of 26 candidate *S. cerevisiae* open reading frames (ORFs).

The techniques used to perform the 1999 search relied heavily on the BLAST algorithm (11), a tool that performs sequence similarity searches. In this work, we describe three extensions of the search protocol for novel methyltransferases. Firstly, we have redefined the motifs using a positionally sensitive scoring matrix, for example where the first letter in the motif might be considered more important for a match than the third letter. Secondly, we have defined these motifs using an assortment of known methyltransferases with different substrate specificities. Finally, we have automated these tasks for easy refinement as more methyltransferases are discovered and to allow for the rapid screening of new genomes as they are sequenced. The results of motif analyses were verified and in some cases extended using sequence profile analysis implemented in PSI-BLAST (12) and HMMer (13), arguably two of the best tools for detection of the remote sequence homology.

From the ‡Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, California 90095-1569 and the §Department of Microbiology, Montana State University, Bozeman, Montana 59717-3520

Received, April 23, 2003, and in revised form, July 17, 2003

Published, MCP Papers in Press, July 18, 2003, DOI 10.1074/mcp.M300037-MCP200

[1] The abbreviations used are: AdoMet, *S*-adenosylmethionine; MSD, methyltransferase-specific database; SM^2, sensitized matrices for scoring methyltransferases; ORF, open reading frame; SGD, *Saccharomyces* Genome Database; SQL, structured query language.

TABLE I
Methyltransferase-specific database

SQL Code[a]	Comments[b]
CREATE TABLE **methyltransferases** (	
methyltransferase_id int(5) DEFAULT '0' NOT NULL auto_increment,	Unique ID for each entry
orf varchar(255) NOT NULL,	*
genename varchar(255) NOT NULL,	*
organism ENUM('cerevisiae', 'pombe', 'coli', 'human') DEFAULT 'cerevisiae' NOT NULL,	Source organism for future use with other species.
majorclass ENUM('unknown', 'nucleic', 'protein', 'lipid', 'small' DEFAULT 'unknown' NOT NULL,	Class of methyltransferase, if known.
beenviewed enum ('Y','N') default 'N' ,	*
mt_verif_status int NOT NULL DEFAULT 0,	*
description text NOT NULL,	Curator annotation
references_pmid text NOT NULL,	PubMed relevant ID's
annotation text NOT NULL,	*
autoid_sets text NOT NULL,	Source of entry into DB
PRIMARY KEY (methyltransferase_id)	
);	

[a] Below is the SQL code used to create the methyltransferase table; variable names are shown in bold type.

[b] Below are descriptions of the fields in the methyltransferase table. Comments of * are described here. **orf** represents the common ORF name for the reading frame examined. **genename** is the common name used to refer to the gene product, if one exists. **beenviewed** reflects that entries can automatically be added to the database; if **beenviewed** is 1 it reflects that the database curator has viewed and commented on a particular record. **mt_verif_status** is a rating (−3–3) of how well the evidence refutes (−3, very refuted) or supports (3, very supported) whether a particular ORF is a methyltransferase; 0 represents no experimental evidence. **annotation** is used to hold the description from an outside database (*e.g.* SGD or Proteome (IncyteGenomics Yeast Proteome Database at www.incyte.com/bioknowledge)). **autoid_sets** is a description of why the entry was added to the database (*e.g.* which programmatic run identified this ORF as a methyltransferase).

EXPERIMENTAL PROCEDURES

Development of a Methyltransferase-specific Database for Automated ORF Tracking and Scoring—To track the progress of automated methyltransferase assignment methodologies, a database of existing yeast methyltransferases was built that could be queried by automated scoring systems. The database system chosen was MySQL (www.mysql.com/), a freely available SQL (structured query language) implementation. The layout of the database, designated MSD for "methyltransferase-specific database," is shown in Table I and is populated as described below.

The MSD is a hand-curated database of methyltransferases combining annotations of genes identified by literature review and genes identified from our automated identification methodologies. Each entry (record) of the MSD is characterized by a number of pieces of information (fields) useful specifically for work with methyltransferases. These include the class of the methyl-accepting substrate, the source organism, the ORF and gene names, and a confidence number that we assigned based on the biochemical evidence in the literature for methyltransferase function. Records are added to the database as new information in the literature becomes available or as candidates are selected based on automatic methyltransferase prediction algorithms. The confidence number in a record runs from 3 (strong experimental support for methyltransferase activity associated with the gene product) to −3 (strong experimental evidence against being a methyltransferase); an entry of 0 denotes no information is available. This MSD is the only database of manually collected and annotated methyltransferases that we are aware of and is available at www.methyltransferase.org/.

In addition to the MSD that we have built, the *Saccharomyces* Genome Database (SGD)[2] provides two databases of gene annotation. We have regularly downloaded these from the SGD and loaded them into local MySQL tables with similar table definitions to those used by SGD. The two databases are available as ftp://genome-ftp.stanford.edu/pub/yeast/tables/ORF_Descriptions/orf_geneontology.tab and ftp://genome-ftp.stanford.edu/pub/yeast/gene_registry/registry.genenames.tab.

One of the advantages to using MySQL is the multitude of programmatic interfaces. Using the methods described below, lists of putative methyltransferases will be generated, which then can be

[2] Dolinski, K., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hong, E. L., Issel-Tarver, L., Sethuraman, A., Theesfeld, C. L., Binkley, G., Lane, C., Schroeder, M., Dong, S., Weng, S., Andrada, R., Botstein, D., and Cherry, J. M., *Saccharomyces* Genome Database at genome-www.stanford.edu/Saccharomyces/.

automatically scored by querying either our MSD or the SGD using locally built programs.

Non-weighted ("Canonical") Degenerate Pattern Searching—The most straightforward method of motif generation and searching is the process of aligning the amino acid sequences of known methyltransferases in the conserved motif regions and making a consensus sequence based on those regions. This is described as a degenerate pattern as each position can possibly be one of several amino acids and as non-weighted because no position is considered more or less important than another (Table II).

To search for these non-weighted degenerate motifs in the translated yeast genome, a FASTA format file containing translations of all of the yeast genomic and mitochondrial genes (from SGD, "orf_trans-.fasta") was modified to remove line breaks within sequences and then searched using the standard UNIX utility, "grep," with appropriate regular expressions describing the degenerate motif (*e.g.* "grep [VIL][ILY][GPY]"). To search for patterns with errors where the position does not match the motif description, either the tool "agrep" (S. W. Wu and U. Manber, ftp://ftp.cs.arizona.edu/agrep/) was used, or in cases with complexity too great to be handled by agrep, a programmed wrapper around grep was used to make repeated queries with a shifting wild card for one or more positions (*e.g.* "grep .[ILY][YPY]; grep [VIL].[GPY]"). A complete schematic for the formulation and use of these non-weighted degenerate motifs is shown in Fig. 1, and a list of the 18 methyltransferases used in the motif definition is shown in Table III.

TABLE II
Description of a degenerate motif pattern

Sample aligned sequences	Combined sequences using a regular expression syntax
..VLDIGPGTG.. ..ALDVGSGPG.. ..ALDLGPGTF..	[AV]LD[IVL]G[PS]G[TP][GF]

Weighted Position-based Motif Searching—The program MEME (17) was used to automatically scan a training set of known methyltransferase amino acid sequences and produce a list of log-odds matrices of amino acids and positions that described putative methyltransferase motifs. These log-odds matrices were then used to scan the *S. cerevisiae* genome using the program MAST (18). The two major obstacles were the formulation of the initial training set and trying to generate motifs that simulated the known variation in spacing between motifs.

The MEME training set was built as follows and is represented graphically in Fig. 2. Entrez (the National Center for Biotechnology Information database query tool, www.ncbi.nlm.nih.gov/) was queried for the keyword "methyltransferase." Of the 5845 matches, all entries not from the RefSeq database were removed; the RefSeq database is the National Center for Biotechnology Information's curated set of entries that are designed to reflect the most highly accurate entries. The remaining 1064 entries were pruned using BLAST such that the final set did not contain any two sequences that matched with an expect value less than the desired threshold using the Blosum62 scoring matrix; the purpose of this culling is to remove entries that are highly similar to one another, which would lead to overrepresentation of certain sequences. With a cutoff expect value of 10^{-20}, 289 sequences were in the final training set, of which 173 contributed to the definition of Motif I; the other 116 did not have regions similar enough to contribute to the Motif I definition and may represent Class II or III methyltransferases.

The output from the MEME program is a list of motifs described as

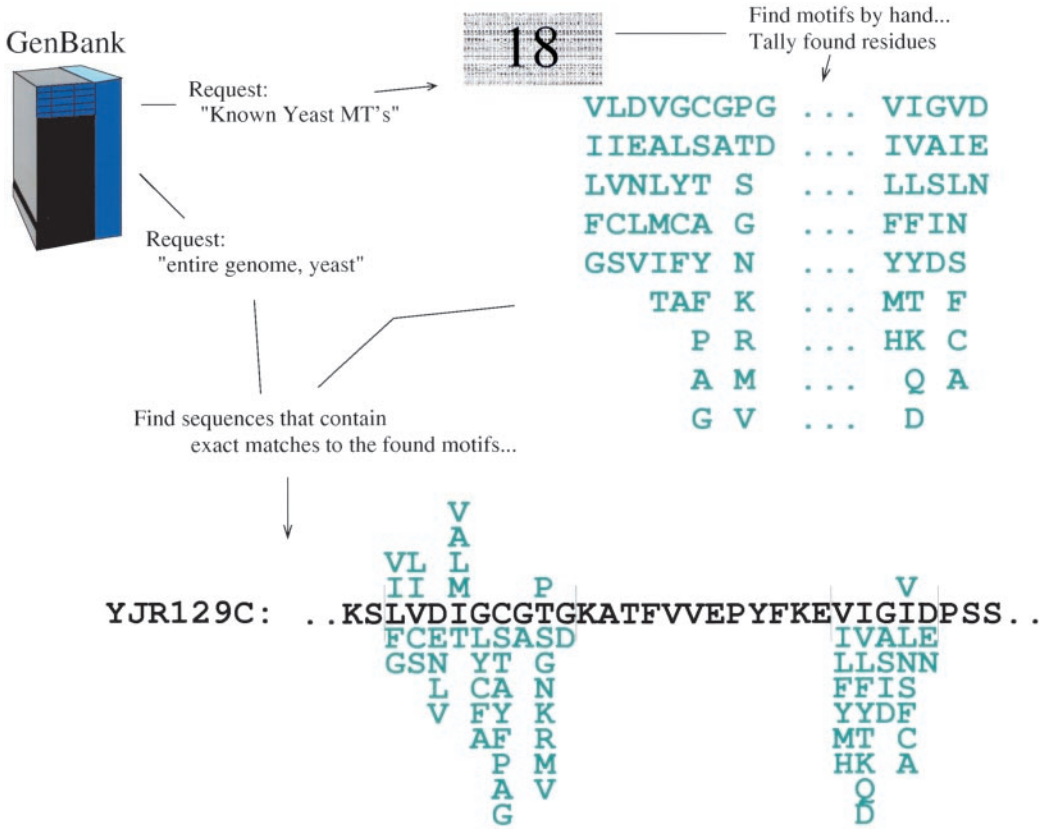

FIG. 1. **Canonical motif formation and use.** *MT*'s, methyltransferases.

TABLE III

Sequence and spacing of motifs in known yeast methyltransferases

Based on literature and database review up to September 2002, 18 highly experimentally supported methyltransferases were selected and visually inspected for the motif sequences. The names of the selections, the motif sequences, and the spacing between them (SP) are shown in the table. Less confident motif assignments are marked with a question mark, and motif identification different than that described previously by Niewmierzycka and Clarke (10) is shown in italics. At the bottom of the table is a graphical representation of the consensus sequence for each of the motifs where the bigger the letter is, the more it occurred in that position in that motif.

ORF	GENE	Motif I	SP	Post I	SP	Motif II	SP	Motif III
YBL024W	TRM4(NCL1)	VLDMCAAPG	20	VVAND	48	FDRIL	42	LLKNNGRLVY
YBR034C	RMT1	VLDVGCGTG	13	VIGVD	37	VDIIL	23	YLVEGGLIFP
YBR133C	HSL7	VVNLGCGSD	14	YVDID	46	CDLND?	18	IVISECLLCY?
YBR236C	ABD1	VLELGCGKG	13	FIGID	48	CDIVS	23	SLKIGGHFFG
YDR120C	TRM1	ILEALSATG	15	VIAND	45	IDLDP?	13	SIEEGGLMLV?
YDR465C	RMT2(F3)	ILVAGAGRG	20	IIAIE	42	IDLCI	31	IPRSYSSYIA
YER175C	TMT1(F9)	LVDVGCGPG	15	IIGSD	45	IDMIT	19	NLRKDGTIAI
YGL136C	MRM2	ILDLGYAPG	15	ILGVD	92	VDVII	52	LLRPLGSFVC
YJL125C	GCD14(F4)	VIEAGTGSG	12	LFSFE	62	LDLPA?	18	GLCCFSPCIE?
YKR056W	TRM2	LVDAYCGSG	12	VIGVE	31	FESID?		
YKR069W	MET1	ISLVGSGPG	12	*IKSAD*	36	*QELLA*	13	*RLKQGDPYIF*
YML008C	ERG6	VLDVGCGVG	13	VIGLN	38	FDKVY	20	VLKPGGTFAV
YML014W	TRM9 (F6)	GIDVGCGNG	10	IIGSD	32	FDFAI	23	KLRQGGQALI
YML110C	COQ5	FIDVAGGSG	20	MDIVD	49	KDIYT	20	VLKPGGIFYC
YMR228W	MTF1	VLDLYPGVG	14	YSLLE	51	NDKFL	28	LYEDFKCKML
YNL063W	HemK	ICDTFTGTG	14	FTAID	40	IDILT	19	KLFEPRLALV?
YOL096C	COQ3	VLDVGCGGG	14	VQGID	36	FDIIT	21	LNPEKGILFI
YPL266W	DIM1	VLEVGPGTG	12	VVAVE	34	FDICI	17	QPRPPRVSIL

matrices with dimensions of (motif length) × 20, each entry of which represents the log-odds for that amino acid occurring at that motif position. A sample motif is shown in Table IV.

Modified Weighted Position-based Motif Searching—Because of the high degeneracy and narrow width of the Post I motif, it could not be automatically identified. At best MEME was found to return a description of a Motif I, some interleaving residues of low significances, and a Post I motif. However, this description is only applicable to a very limited number of methyltransferases. Instead a matrix describing Post I was hand-forged based on the amino acid frequencies of the Post I motifs in the known *S. cerevisiae* methyltransferases (as of September 2002) as shown in Table III. Additionally, MAST, the

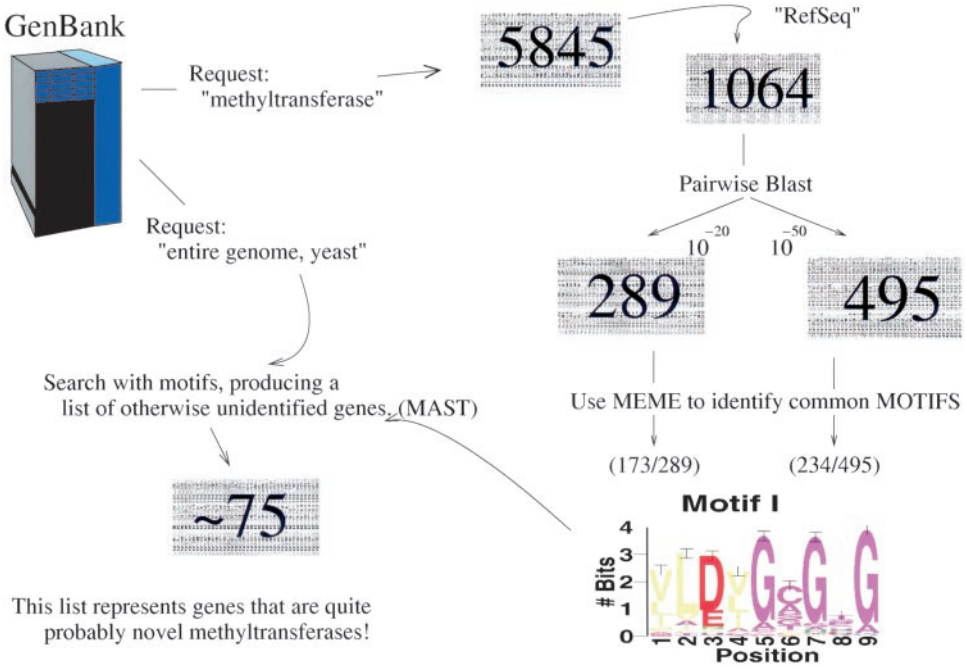

FIG. 2. **MEME training set formulation and automated motif generation.**

tool that searches genomes based on the MEME motifs, does not allow searching for two motifs separated by a variable gap size. Therefore a series of matrices was built with the MEME-determined definition of Motif I and the hand-built definition of Post I separated by between 10 and 35 score-neutral entries in the matrix. Because MAST will, if possible, match multiple motifs to a target sequence (a situation almost guaranteed by the degenerate description of Post I), the source code to MAST needed to be modified to only consider the best fitting motif to any given target.

Automated Scoring of Candidates—Automated methyltransferase identification methods produce lists of gene names (open reading frames) that are putatively methyltransferases. The results of these searches need to be evaluated, including the rather large lists generated by the non-weighted degenerate searches. Evaluation is the process of taking a list of generated candidates and deciding for each candidate whether it is a known methyltransferase (a "hit"), whether it is known not to be a methyltransferase (a "miss" or "false positive"), or whether it is neither of the above (a "putative methyltransferase"). Systematic evaluation is performed as follows. If the candidate is in the MSD, assignment is based on that score (2 or 3 is a hit, −2 or −3 is a miss, and −1, 0, or 1 is a putative methyltransferase). Otherwise the annotation of the two SGDs (orf_geneontology.tab and registry-.genenames.tab) is queried. If the annotations are marked as "unknown" the candidate is considered a putative methyltransferase. If the annotations contain the word "methyltransferase" the candidate is considered a methyltransferase. Otherwise, the candidate is considered an incorrect prediction (a false positive).

There are a number of inconsistencies in the SGD that can lead to inaccurate scoring. For example, HSL7, GCD14, and HemK are still not annotated as methyltransferases in the SGD (although they are in the MSD). This reflects that some genes are annotated as part of a pathway or have a phenotype but that the role as a methyltransferase was not initially known; for example HSL7 (YBR133c) is annotated as a negative regulator of the SWE1 kinase, but experimental evidence has confirmed the prediction of HSL7 as a methyltransferase (19).

Profile Searches Using PSI-BLAST and HMMer—A compilation of protein sequences in SCOP 1.61 (astral.stanford.edu/) and non-redundant SwissProt and TrEMBL databases (ftp://us.expasy.org/databases/sp_tr_nrdb/fasta/) was iteratively searched using the PSI-BLAST program (12). Each potential methyltransferase ORF sequence was used as the query with a profile inclusion E-value threshold of 0.001 and composition-based statistics turned on (20). The iterations were carried out for five rounds (or until convergence), and PSI-BLAST checkpoint files were saved for future use. The results of searches were inspected after each iteration to ensure that no compositionally biased sequences or spurious matches were included in the profile. To increase the sensitivity in the second step, candidate sequences and their corresponding checkpoint files from the first step were used as inputs for PSI-BLAST to scan the yeast proteome (genome-www.stanford.edu/Saccharomyces/). The searches were done for one iteration with the E-value set at 1e-5 to account for the smaller size of the yeast proteome compared with the database used to construct the profile. Potential methyltransferase ORF sequences were also individually compared with the Pfam 8.0 database (pfam.wustl.edu/), a collection of profile-hidden Markov models built from manually curated alignments of more than 5000 protein families (21). The searches employed the *hmmpfam* module of HMMer (13) (hmmer.wustl.edu/), and E-value threshold was set at 1.

RESULTS

Canonical Pattern Searching Markedly Loses Discrimination with Increasing Sensitivity and Does Not Rank Results—The 18 known Class I AdoMet-dependent yeast methyltransferases, based on literature review and database annotation at the time the search was performed, were used to build a set of consensus sequences for the various motifs as shown in Table III. Non-weighted motifs are represented by a regular expression. For example "[VILFG][LIVCS][DE]" specifies that the first position of the sequence must contain a Val, Ile, Leu, Phe, or Gly and that the second position must contain a Leu, Ile, Val, Cys, or Ser and so forth through the pattern. The full

TABLE IV

MEME position-specific log-odds description of Motif I

The 287-sequence training set of known methyltransferases described herein was used as a training set to the MEME program. MEME produces a log-odds matrix (shown below) with one line/position in the motif. Each line has 20 entries, one for each of the amino acids; the order of the amino acids is ACDEFGHIKLMNPQRSTVWY. A key of amino acid positions has been added above the matrix. In the description of Motif I shown below, the most predominant amino acid score for each position is in bold. The most predominant sequence as described in this motif is VLDVGCGTG.

Amino acid residue

A C D E F G H I K L

M N P Q R S T V W Y

Position	Values									
1	-171	75	-735	-668	-15	-249	-549	204	-634	10
	75	-593	-651	-552	-630	-534	-181	**283**	-9	-235
2	-44	55	-696	-620	-101	-682	-539	-22	-597	**287**
	-10	-604	-594	-477	-566	-340	-287	-9	-122	-71
3	-293	-489	**371**	106	-325	-231	-126	-269	-527	-580
	-221	-66	-586	-435	-537	-322	-475	-234	-511	-504
4	-79	44	-736	-669	92	-640	-186	170	-635	129
	20	-593	72	-552	-631	-232	-441	**208**	-526	-235
5	33	83	-516	-588	91	**341**	-548	-682	-569	-731
	-198	-258	-310	-564	-574	-187	-211	-632	-592	-113
6	147	**442**	-327	-674	-130	13	-555	-260	-640	-444
	-70	-597	-106	-235	-636	97	105	-104	-532	-516
7	-155	-170	-501	-573	-685	**377**	-534	-675	-554	-415
	-579	-178	-616	-552	-560	-189	-551	-627	-579	-265
8	-34	120	-159	-109	-194	-151	-30	-19	-97	-148
	-81	4	61	-185	-211	141	**260**	-39	-576	-54
9	-131	-550	-143	-347	-684	**374**	-533	-674	-335	-724
	-579	-436	-616	-551	-559	-237	-159	-627	-578	-625

TABLE V

Non-weighted degenerate search results for yeast methyltransferase motifs

A translated database of yeast genomic and mitochondrial genes ($n = 6312$) was searched for exact matches to canonically defined motifs. "Search set" (sections **A** and **B**) lists the motifs used for a given search as defined in section **C**. The notation of [10..30] reflects that between 10 and 30 amino acids must be between the two flanking motifs. "No. of errors" refers to how many deviations from the described motif are allowed in that search run. "Correct identifications," "Incorrect identifications," and "Unknown" break down the results into those that are methyltransferases, are not methyltransferases, and are of unknown methyltransferase status, respectively. "% Correct" is (number of correct/(number of correct + number of incorrect)) × 100. All results are based on a search of MSD followed by searches of SGD if the MSD search was not productive.

	No. of errors allowed	No. of results	Correct (known) identifications	Incorrect identifications	Known entries correct	Unknown
					%	
A Complete search set						
Motif I	0	62	21	23	48	18
	1	1094	29	610	5	455
Motif I [10..30]						
Post I	0	30	21	2	91	7
	1	128	24	67	26	37
	2	1184	29	644	4	511
B Restricted search set						
Rmotif I	0	11	5	0	100	6
	1	64	11	34	24	19
	2	1014	27	601	4	386
Rmotif I						
[10..30] rpost I	0	7	4	0	100	3
	1	17	11	0	100	6
	2	86	20	35	36	31
	3	591	23	339	6	229
	4	2894	38	1525	2	1331
C Set name	Set description					
Motif I	`[VILFG][LIVCS][DENLV][VALMIT][GLYCFA][CSTAYFPAG][GA][PTSGNKRMV][GD]`					
post I	`[VILFYMH][IVLFYTKQD][GASID][VILNSFCA][DEN]`					
Rmotif I	`[VIL][LIV][DEN][VALMIT][G][CST][G][PTSG][GD]`					
Rpost I	`[VILFYM][IVLFYT][GASID][VILNSFCA][DEN]`					

specification of Motif I and Post I is shown in Table V, part C.

The results of searching with these patterns are shown in Table V. A first search of the yeast genome with Motif I returned 62 ORFs, including 21 known methyltransferases, 23 false positives, and 18 unknowns. When searched with the Motif I-Post I set, 30 ORFs were found, including 21 known methyltransferases, 2 false positives, and 7 unknowns. In the latter analysis, the number of false positives was dramatically reduced, but the number of putative methyltransferases was also much smaller.

In addition to simple searches with the listed patterns, the sensitivity was increased by allowing errors (deviations from the proscribed pattern) to be introduced. As shown in Table V, part A, the number of results grows quickly as multiple deviations are allowed. However, the number of false positives (candidates that have a known non-methyltransferase function) also increases rapidly as deviations are allowed, suggesting that this approach is not a good one for identifying new methyltransferases. The large number of false positives comes from the fact that a best match at each position is accepted just as readily as a worst match at each position. For example, VLDVGCGPG is treated no differently than GSVTAAAVD; the latter would not be considered an acceptable Motif I based on known methyltransferase sequences.

In an attempt to reduce false positives, a restricted search motif was created by removing the unusual amino acids from the patterns (Table V, part C). The results from searching with the restricted motif sets are shown in Table V, part B. Although the initial number of matches is lower, the amount of information returned is similar (for a given number of results, the partitioning of the results into "correct," "incorrect," and unknown is similar to that seen in Table V, part A). It is clear from these results that there is a very low limit of information that can be derived from these types of canonical searches before the signal-to-noise ratio drops well below an acceptable limit.

Unsupervised Automatic Motif-based Searches Are Similar to Human-mediated BLAST Searches and Can Be Greatly Improved with Minor Parameter Modification—We then took a second approach to finding new methyltransferases using automated motif identification processes. To answer the question as to how good default "out-of-the-box" motif identification and searching is, the 1064 RefSeq matches for the

TABLE VI
Scored automated motif identifications

The results of the 1999 yeast methyltransferase search (10) are compared to a subset of the weighted position-specific motif ("MEME") searches. MAST results are listed (in order of motif match significance) until and including the fifth known false positive. Results are broken down into positives ("Correct ID's"), false positives ("Incorrect ID's"), and putatives ("Unknown"). % Correct is (Correct/(Correct + Incorrect) × 100). Result order is shown graphically with the X's being positives, Y's being negatives, 0's being unknowns, and u's being SGD-identified but without the keyword "methyltransferase."

Set Identification	Correct ID's	Incorrect ID's	% Correct	Unknown	
1999:	3	3	50%	18	
MEME, expect 10^{-20}, all Motifs:	9	5	64%	17	
MEME, expect 10^{-20}, Motif I:	13	5	72%	26	
MEME, expect 10^{-50}, Motif I:	12	5	71%	24	
MEME, expect 10^{-20}, Motif I-[10-30]-Post I:	14	5	74%	27	
MEME, expect 10^{-20}, Motif I-[5-25]-Post I:	14	5	74%	27	
MEME, expect 10^{-50}, Motif I-[5-25]-Post I:	16	5	76%	28	
MEME, expect 10^{-50}, Motif I-[10-30]-Post I:	16	5	76%	28	

keyword search "methyltransferase" were BLASTed against themselves to return sets in which no entry was homologous to any other entry with a significance greater than a certain expect value. The two expect values used were 10^{-20} and 10^{-50}, which returned training sets of 289 and 495 sequences, respectively. The motif-searching program MEME (17) was trained with the 10^{-20} set without parameter modification and used in the default mode to detect five motifs. The matrices obtained were then used by the MAST program (18) to search a yeast-translated ORF database for matches (Table VI, MEME expect 10^{-20}, all motifs). Here, 9 methyltransferases were returned, with 5 false positives and 17 unknowns.

During inspection of the generated motifs, it was noted that MEME was generating motifs that were not specific to all Class I AdoMet-dependent methyltransferases; for example the NPPY motif common to only the DNA N^6-adenine methyltransferases and protein glutamine methyltransferases (22–24) was found. The searches were thus repeated using only the motifs returned that were similar to the already known methyltransferase motifs and again only with the automatically generated Motif I. This modification resulted in an improved performance with 13 correct methyltransferases returned along with 5 false positives and 26 unknowns (Table VI, MEME expect 10^{-20}, Motif I).

Use of the automated MEME-MAST tool set in its default configuration was able to create lists of putative methyltransferases that were similar to those that were obtained by hand-using BLAST and manual sequence inspection (10). The advantage in using the automated tools is that they involve less effort and can therefore be rapidly applied to other genomes. Additionally, the results returned by the MEME-MAST tool set were significantly improved over the manual method by performing a first pass analysis of the results and rerunning the search after removing the non-Motif I confounding elements that were specific to only certain subclasses of methyltransferases or that may represent motifs for distinct types of enzymes such as the related NAD/NADP dehydrogenases.

It is worthwhile to note that a major difference between this method and the canonical method described above is that this method begins with a list of gene sequences, which are then ordered in terms of likelihood of each entry being a methyltransferase. Using default settings, only the top percentage of entries is returned. However, with reduced reporting stringency, the entire genome can be ordered by the likelihood of each ORF being a methyltransferase.

A Less Stringent Training Set Produces Slightly Improved Results When Combined with a Variably Distal Hand-coded Post I Motif: Sensitized Methyltransferase-scoring Matrices (SM^2)—The MAST program returns a score on every hit that represents how well the subsequence of the ORF (motif) fits the MEME-derived scoring matrix. However, this score is not directly a probability of the gene product of a sequence being a methyltransferase. To compare sets of results returned from MAST, which vary in both order and motif match significance scores, we have arbitrarily chosen a cutoff point at the fifth known incorrect identification. Comparing the results of the 10^{-20} and 10^{-50} training sets yields very similar results (Table VI; MEME expect 10^{-20}, Motif I and MEME expect 10^{-50}, Motif I) with the 10^{-20} results being slightly better (one additional positive match and two additional candidates); this is the more stringent of the two sets.

Noting the highly conserved, albeit degenerate, Post I motif, a set of hand-coded matrices describing the Post I motif was appended to the description of Motif I in an attempt to improve the search sensitivity. The set varied only in the number of score-neutral elements that separated the Motif I and Post I motifs. Two spacings considered were 5–25 and 10–30. The results are shown in Table VI (MEME, expect 10^{-20}, Motif

TABLE VII
Ordered output of S. cerevisiae putative Class I methyltransferase ORFs[a]

Cumulative percent correct based on genes of known functions	ORF name and common names for genes of unknown function	ORF name and common names for genes with known function Known methyltransferases are shown in left–justified GREEN CAPITALS; known non–methyltransferases are shown in right–justified *red italics*			
	Names in **bold** were identified as putative methyltransferases F1–F26 (10)		PSI-BLAST NR[b]	PSI-BLAST Profile SGD[c]	PFAM ID[d]
Percent Correct	Putative ORF's	Known ORF's			
		YBR034C RMT1 HMT1	+	+	+
		YOL096C COQ3	+	+	+/–
		YML008C ERG6	+	+	+
100	ynl024c –		+	+	–
100	**yhr209w (F10)**		+	+	+
			+	+	+
100	**yjr129c (F20)**		+	+	–
100	**ycr047c (F7)** bud23		+	+	+/–
		YER175C (F9) TMT1	+	+	+/–
		YBR236C ABD1	+	+	–
100	**ybr271w (F16)**		+	+	–
100	**yor240w/239w (F12)** abp140		+	+	+/–
100		**YDL201W (F1)** TRM8	+	+	+
		YPL266W DIM1	+	+	+
100	**ydr140w (F8)**		+	+	–
100	**yil064w (F11)**		+	+	–
100	**ydr316w (F2)**		+	+	+
		YJL125C (F4) GCD14	+	+	–
100	ypl017c –		–	–	–
		YBR061C TRM7	+	+	+
		YML014W (F6) TRM9	+	+	–
100		YHR070W TRM5	+	+	–
		YBL024W TRM4 NCL1	+	+	+
100	**ynl092w (F23)**		+	+	–
100	ylr146c spe4[e]		+	+	+
100	**yil110w (F18)**		+	+	–
		YML110C COQ5	+	+	+
100	ynl061w nop2		+	+	+
100	ymr228w mtf1		+	+	–
100	yol124c –		+	+	+
		YGL136C MRM2	+	+	+
		YDR440W DOT1 PCH1	+	+	–
		yer043c sah1	–	–	–

I-[10–30]-Post I; MEME, expect 10^{-20}, Motif I-[5–25]-Post I; MEME, expect 10^{-50}, Motif I-[5–25]-Post I; and MEME, expect 10^{-50}, Motif I-[10–30]-Post I).

The results for all four sets are quite similar to one another and slightly improved over the non-Post I searches (14–16 correct identifications and 27–28 candidates). Although the ordering of the ORFs was different, the significance of the results was similar based on the number of correct identifications and number of candidates returned for the 5–25 and 10–30 spacing. The 10^{-50} training set returned slightly better results than the 10^{-20} training set with two additional correct identifications and one additional candidate ORF.

TABLE VII—*continued*

		YNL063W HEMK PRMC	+	+	–
		yk1085w mdh1 acn50	–	–	–
		YCL054W SPB1	+	+	+
90	**ybr261c (F15)**		+	+	–
		yor375c gdh1	–	–	–
		yal062w gdh3	–	–	–
		YDR435C PPM1	+	+	+
83	ypr069c spe3[e]		+	+	+
		YDR465C (F3) RMT2	+	+	–
83	yml005w –		+	+	–
		YKR056W TRM2 RNC1	+	+	–
84	ygr155w –		–	–	–
		YDL014W (F13) NOP1 LOT3	+	+	+
85	ypl074w yta6		–	–	–
85	ygl004c –		–	–	–
		yil160c pot1	–	–	–
		q0045 (F25) *cox1*	–	–	–
		q0065 ai4	–	–	–
		yhr176w	–	–	–
73	ylr063w –		+	–	–
		yol152w fre7	–	–	–
		ynl183c npr1	–	–	–
69	ynr060w fre4		–	–	–
		q0070 ai5_alpha	–	–	–
		yer089c ptc2	–	–	–
65	ygr012w –		–	–	–
		ymr303c adh2	–	–	–
63	**ydr083w (F17)** rrp8		+	+	+
63	**ylr285w (F22)** nnt1[f]		+	+	–
63	yhr092c hxt4		–	–	–
		ydr345c hxt3	–	–	–
		ydr343c hxt6	–	–	–
59	ydr342c hxt7		–	–	–
59	yil170w hxt12		–	–	–
59	ylr107w rex3		–	–	–
		ygl044c rna15	–	–	–
58	ynr062c –		–	–	–
		yel042w gda1	–	–	–
		ygr244c lsc2	–	–	–
55	yol156w hxt11		–	–	–
55	yjl219w hxt9		–	–	–
		ybl056w ptc3	–	–	–
54	ymr244w –		–	–	–
54	ypr080w –		–	–	–
		ybr118w tef2	–	–	–
52	ydl168w –		–	–	–
52	**ynr029c (F24)**		–	–	–
52	ykl155c rsm22 rsm50		+	–	–
		yer095w (F26) *rad51*	–	–	–
		yhr042w ncp1	–	–	–
		yhr096c hxt5	–	–	–
49	ykl150w mcr1		–	–	–
		ybr183w ypc1	–	–	–
		yc1018w leu2	–	–	–
		ylr260w lcb5	–	–	–
46	yal046c –		–	–	–

We describe this optimized scoring system as sensitized matrices for scoring methyltransferases (SM^2). The results from the best training set are expanded in Table VII, which represents our new best list of putative methyltransferases in yeast. Descriptions of all the currently known *S. cerevisiae* methyltransferases are shown in Table VIII.

TABLE VII—*continued*

		yor202w his3	–	–	–
45	ynl151c rpc31		–	–	–
		ygr255c cog6	–	–	–
		ymr154c rim13	–	–	–
43	ypl030w –		–	–	–
43	yol141w ppm2		–	–	–
		yhr094c hxt1	–	–	–
42	ymr116c asc1		–	–	–
42	yjl146w ids2		–	–	–
42	ydl062w –		–	–	–
		yol086c adh1	–	–	–
		yer171w rad3	–	–	–
	YKR069W MET1		+	+	+(*)
		ypl252c yah1	–	–	–
	YDR120C TRM1[g]		+	+	+
	ynl022c –[g]		+	+	+
	ybr141c –[g]		+	+	–

[a] This table represents our current best list of putative methyltransferases in order of SM^2 significance. ORFs are listed in two columns; the "Putative ORF" column is the list of ORF names (and, if available, common names) of unknown function, and the "Known ORF" column shows ORF names of known function. Within the known ORF column, green entries in capital letters on the left are experimentally confirmed methyltransferases, and red entries in italics on the right are proteins with identified non-methyltransferase function(s). "Cumulative Percent Correct" is based on the correct and incorrect matches in the known ORF column. All ORFs identified in 1999 (shown in bold type with the previous F-designation in parentheses) (10) are identified here, except for those that fell below the significance cutoff for the table: yjr072c (F19), ylr137w (F21), and the known false positive yal061w (F14; FUN50).

[b] A plus is recorded if a PSI-BLAST search of the putative methyltransferase entry against the non-redundant protein database as described under "Experimental Procedures" recovers any AdoMet-dependent methyltransferase with an E-value <0.001.

[c] A plus is recorded if a search of the SGD with the profiles generated in the PSI-BLAST search matches a known methyltransferase with an E-value $<$1e-5.

[d] A plus is recorded if a search of the Pfam 8.0 database recovers any methyltransferase with an E-value <0.1. A plus/minus is recorded if $0.1 <$ E-value <1.0.

[e] spe3 and spe4 encode spermidine and spermine synthases, respectively. The encoded amino acid sequences are very similar to those of plant putrescine *N*-methyltransferases, but no methyltransferase activity of the yeast proteins has been shown.

[f] Indirect evidence has been presented for the function of the ylr285w gene product as a nicotinamide *N*-methyltransferase (25).

[g] These ORFs, although below the inclusion threshold of the rest of the table entries, are included because they appear with high significance in the PSI-BLAST analysis.

Sequence Profile Analysis with PSI-BLAST and HMMer Provides Support for SM^2 Results—To test the validity of the SM^2 analysis, all ORFs listed in Table VII were probed individually using PSI-BLAST and HMMer, two powerful profile-based search tools that have been used in recent years with great success to detect remote sequence homology. Each sequence was first searched with PSI-BLAST against the non-redundant protein database in an attempt to provide support for its inclusion into the methyltransferase superfamily. Those candidates that matched known methyltransferases at E-value <0.001 before the sequence in question was included in the profile were considered true positives. Here, all true positives matched numerous methyltransferases, sometimes even in the first iteration. The subsequent iterations were important in generating checkpoint files, which correspond to position-specific scoring matrices. The checkpoint files were then used to increase the sensitivity of search against the yeast proteome. We annotated as true positives all sequences that identified a known methyltransferase in the yeast proteome at an E-value $<$1e-5, or those that were recovered themselves by another query using the same E-value threshold.

As can be seen in Table VII, most candidates with percent correct values of 80 or greater pass as true positives according to PSI-BLAST criteria. Therefore, it appears that percent correct value 80 can be used in most cases as a safe threshold for automatic functional assignments. However, this analysis also showed that two candidates with lower percent correct values (YDR083w and YLR285w) are likely to be true positives, cautioning against the strict threshold. Finally, three ORFs not originally included in Table VII (YDR120c, YNL022c, and YBR141c) were identified as potential methyltransferases

TABLE VIII

AdoMet-dependent methyltransferases in S. cerevisiae

This table lists all of the currently identified *S. cerevisiae* methyltransferases. Genes marked with "*" are genes that are not listed in Table VII; HSL7 is found in the 38th cumulative percentile, and TRM1 is not found through the 30th cumulative percentile (this is expected considering its very unusual Motif I, "ILEALSATG," Table III). The entry marked with "1" is for MTF1; although there is no enzymatic evidence for this entry being a methyltransferase, the crystal structure is very similar to other known AdoMet-dependent methyltransferase structures (14).

Gene identified, seven beta-strand motifs are present	**EC #**	**Common Name**	**ORF**
RNA 2'-O-MT		NOP1	YDL014W
mitochondrial 21 S rRNA 2'-O-uridine-2791 O-MT	2.1.1.57	MRM2	YGL136C
60 S rRNA MT		SPB1	YCL054W
mRNA cap N-7-guaninosine N-MT	2.1.1.56	ABD1	YBR236C
rRNA 18S N6,N6-dimethyladenine N-MT	2.1.1.48	DIM1	YPL266W
tRNA N2,N2-guanosine-26 N-MT	2.1.1.32	TRM1	YDR120C*
tRNA 5-uridine-54 C-MT	2.1.1.35	TRM2	YKR056W
tRNA 5-cytosine C-MT	2.1.1.29	TRM4	YBL024W
tRNA 1-guanosine-37 N-MT	2.1.1.31	TRM5	YHR070W
tRNA 1-adenosine-58 N-MT	2.1.1.36	GCD14	YJL125C
tRNA 2'-ribose-32,34-O-MT		TRM7	YBR061C
tRNA 7-guanosine N-MT	2.1.1.33	TRM8	YDL201W
tRNA 5-carboxymethyluridine O-MT		TRM9	YML014W
snRNA and snoRNA 5'-cap NH2-2,2'-guanine N-MT		TGS1	YPL157w
RNA MT?		MTF1	YMR228w[1]
Histone H3 protein lysine-79 MT		DOT1	YDR440W
protein omega-arginine MT type I (asymmetric di)	2.1.1.125	HMT1/RMT1	YBR034C
protein delta-arginine MT (ribosomal L12 protein MT)		RMT2	YDR465C
protein omega-arginine MT type II	2.1.1.126	HSL7	YBR133C*
PP2A protein carboxyl O-MT (C-terminal Leu)		PPM1	YDR435C
protein glutamine N-MT (Release factor; HemK homolog)		HEMK/PRMC	YNL063W
uroporphyrin-III C-MT	2.1.1.107	MET1	YKR069W
trans-aconitate O-MT	2.1.1.145	TMT1	YER175C
Delta-24-sterol C-MT	2.1.1.41	ERG6	YML008C
3,4,dihydroxyl-5-hexaprenylbenzoate O-MT	2.1.1.64	COQ3	YOL096C
COQ5 C-MT		COQ5	YML110C
Gene has been identified, seven beta-strand motifs are not present			
mRNA 6-adenosine N-MT	2.1.1.62	IME4/SPO8	YGL192W
Mitochondrial 21S rRNA 2'-O-guanosine -2270 O-MT	2.1.1.51	PET56	YOR201C
tRNA 2'-O-guanosine-18 O-MT	2.1.1.34	TRM3	YDL112W
tRNA 1-guanosine-9 N-MT		TRM10	YOL093W
Cytochrome c trimethyllysine N-MT	2.1.1.59	CTM1	YHR109W
Histone H-3 Lysine-4 N-MT	2.1.1.43	SET1	YHR119W
Histone H-3 Lysine-36 N-MT	2.1.1.43	SET2	YJL168C
Isoprenylcysteine O-MT	2.1.1.100	STE14	YDR410C
Diphthamide N-MT	2.1.1.98	DPH5	YLR172C
Phosphatidylethanolamine N-MT	2.1.1.17	PEM1	YGR157W
Phospholipid N-MT	2.1.1.71	PEM2	YJR073C
AdoMet-homocysteine S-MT	2.1.1.10	SAM4	YPL273W

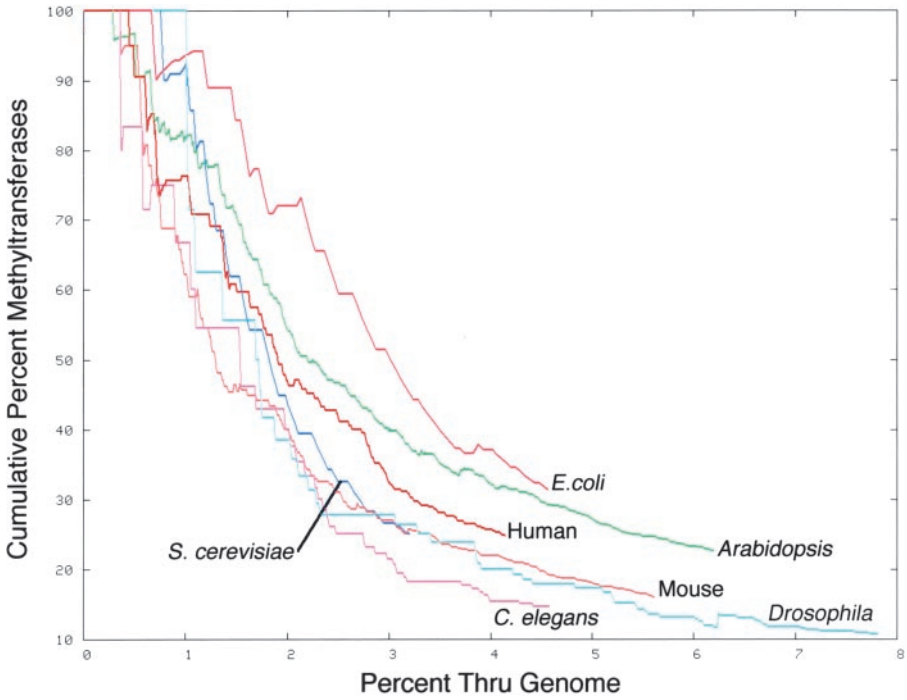

FIG. 3. **Success of putative Class I (seven β-strand) methyltransferase identification in seven genomes based upon the cumulative percent methyltransferases of genes of known function.** Functionally annotated ORFs of seven genomes were ordered based upon their likelihood of being methyltransferases using the SM^2 method. For each annotated ORF, we plot the cumulative percent methyltransferases (MT's) found up to that point (number of MT's/(number of MT's + number of non-MT's) $\times$ 100) against its SM^2 ordered position in the genome (ordered position/total genes $\times$ 100).

because other queries matched them at an E-value <1e-5. These proteins were subsequently used as queries with the non-redundant protein database and fulfilled the criteria outlined above for inclusion in the methyltransferase superfamily.

Sequence comparisons with HMMer tools and the Pfam 8.0 database provided further support for slightly more than half of PSI-BLAST true positives but were ultimately less informative than the SM^2 method described here despite the fact that Pfam 8.0 contains HMMs for more than 30 methyltransferase families, including some families that are presently annotated as uncharacterized.[3] Although it is formally possible that some true positives from the SM^2 and PSI-BLAST searches represent false predictions and as such were not confirmed by HMMer, it is clear that the coverage of the methyltransferases superfamily in Pfam 8.0 is far from reaching saturation.

SM^2 Methodologies Are Easily Applied to Other Genomes and Show Results Similar to Those Seen in S. cerevisiae—To generalize these results, translated ORFs from six additional recently sequenced genomes (human, mouse, *Drosophila*, *Caenorhabditis elegans*, *Arabidopsis*, and *Escherichia coli*) were ordered based on likelihood of being a methyltransferase using the MAST tool in the SM^2 configuration with the "expect 10^{-50}, Motif I-[10–30]-Post I" criterion described in Table VI. Lists of putative methyltransferases generated by this method are in the on-line supplement to this paper.

The methods developed here appear to have similar success in finding methyltransferases in these other genomes. The efficacy of ordering the genome in terms of likelihood of being a methyltransferase is shown graphically in Figs. 3 and 4. After the genome is ordered in this fashion, one can look at the genes of known function and develop an overall cumulative percent methyltransferases expression that is similar to the scoring methodology used earlier in this article and shown graphically in Fig. 3. A possibly more telling view is to, at each point in the genome, look at the local percent methyltransferases, that is, what is the percent methyltransferases in a small window surrounding the position we are looking at. Fig. 4 shows this graphically using a window size of 0.01% of the genome size. As can be seen, the percent likelihood of finding a methyltransferase rapidly falls off after the top scores in 2–3% of the genome are analyzed.

The main difference between the scoring used here and the scoring used earlier in this paper with the yeast genome is that the MSD was used there to confirm the assignment of function. Here, the shortcut of looking solely at the provided gene annotation is used.

The final calculation in this section is the prediction of the total number of motif-bearing methyltransferases in a given

[3] M. Dlakić, unpublished results.

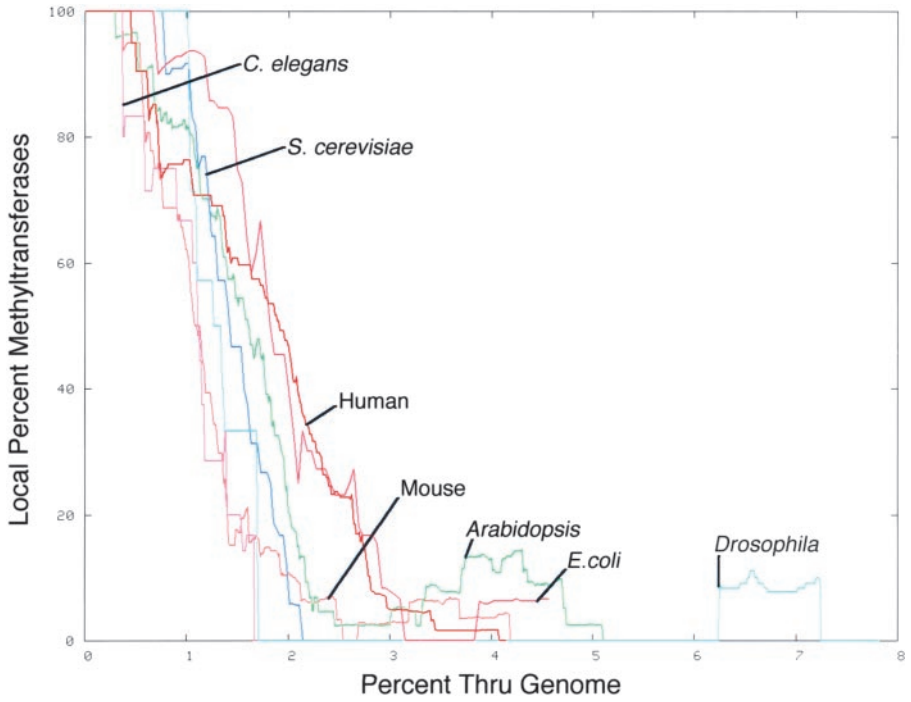

FIG. 4. **Success of putative Class I methyltransferase identification based upon the local percent methyltransferases of genes of known function.** Functionally annotated ORFs of seven genomes were ordered based upon their likelihood of being methyltransferases using the SM^2 method. For each annotated ORF, we plot the percent methyltransferases (MT's) found in a window of 0.01% of the genome (number of MT's/(number of MT's + number of non-MT's) × 100) against its SM^2 ordered position in the genome (ordered position/total genes × 100).

genome. This calculation was done by taking the data from Fig. 3 and performing the following calculation for each point

$$\%_{in\ genome} = 100 \times \frac{\frac{\%_{correct}}{100} \times \left(\frac{\%_{thru\ genome}}{100} \times Number\ Genes\ In\ Genome \right)}{Number\ Genes\ In\ Genome} \quad \text{(Eq. 1)}$$

or simply,

$$\%_{in\ genome} = \frac{\%_{correct} \times \%_{thru\ genome}}{100} \quad \text{(Eq. 2)}$$

This data is plotted in Fig. 5. It is predicted from the graph that all the genomes assayed have a similar percentage (0.6–1.6%) of genes that are of the Class I motif form of methyltransferases.

DISCUSSION

The purpose of this study was to identify novel methyltransferases using the primary sequence data available from genome sequencing projects. We have developed semi-automated methods that order the encoded amino acid sequences of the open reading frames of a genome in terms of their likelihood of being Class I methyltransferases (seven β-strand family). Using the criteria of getting as many of the known methyltransferases in our list as possible while, at the same time, keeping the number of known false positives to a minimum, we have identified candidate methyltransferases in yeast and other organisms. This system is automated enough to be easily applicable to new genomes as they are sequenced. It is also easy to recompute the training set as additional validated methyltransferases become known, allowing for the generation of updated candidate lists.

Including an ORF in a list of putative methyltransferases is obviously only a first step toward biochemically characterizing a new AdoMet-dependent methyltransferase. Even if we had a perfect method that identified all the AdoMet-dependent genes in a genome, we would still need to determine what their methyl-accepting substrates were to define their biological function. As enzymatic activity specification is the slow step in this process, it is sufficient at this point to have a partial list with even marginal confidence that each entry in the list is a methyltransferase. Having a list of 100 ORFs where each entry is 50% likely to be a methyltransferase is much better than having an entire genome ORF list where each entry is only 1–2% likely to be a methyltransferase. As time progresses and these early lists are exhausted, better techniques will hopefully evolve for protein identification that will allow establishing a complete catalog of the methyltransferase complement of an organism.

In the end, only time will tell if we have, in fact, generated here "good" lists of candidate methyltransferases. We can say at this point, however, that our methodology does appear to

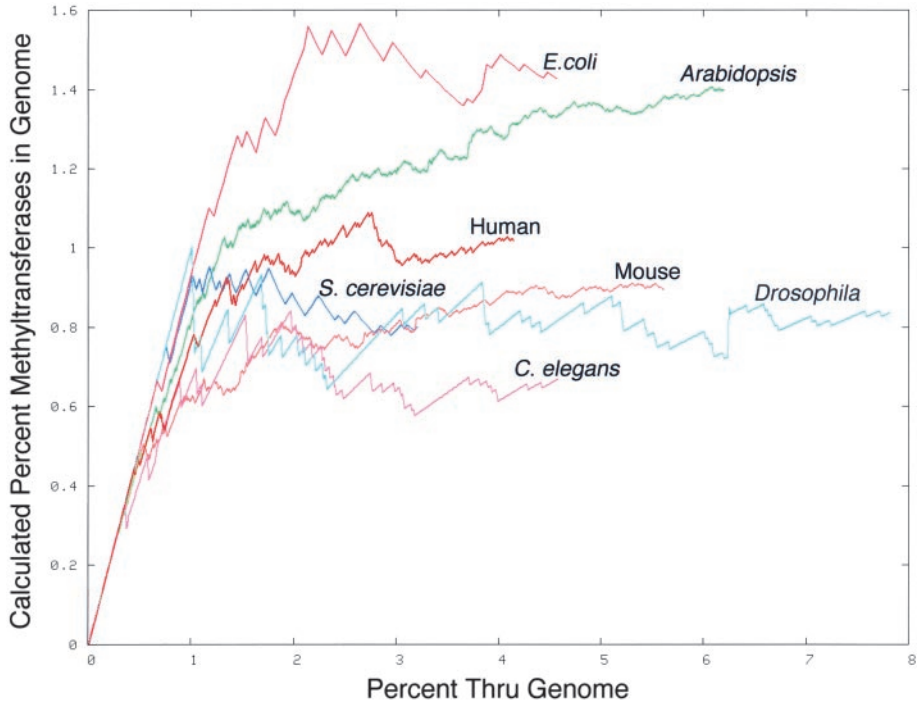

FIG. 5. **Calculation of the percent of the genome comprising Class I methyltransferases based upon identified genes of known function in seven genomes.** Functionally annotated ORFs of seven genomes were ordered based upon their likelihood of being methyltransferases using the SM^2 method. For each annotated ORF, we plot the predicted number of methyltransferases in the genome to that scoring value as described in Equation 2.

be superior to that presently employed in a database such as Pfam (21). For example, of the 24 experimentally verified yeast methyltransferases described in Table VII, eight are not annotated as methyltransferases in version 8.0 of the Pfam database. Additionally, we note that the SM^2 methodology used here has identified six new candidates in the "100%" region and 33 new candidates in the "100–42%" region of Table VII that were not detected in the 1999 analysis of yeast proteins (10). We have been pleased to see a steady progression of our best yeast candidates into the class of experimentally supported methyltransferases. For example, just in the time between the completion of this manuscript and its revision, two of our high scoring candidates were identified as specific methyltransferases (15, 16). Further evidence of this progress is that in 1999 only seven Class I methyltransferases had been described in yeast (10); the present number is 26 (Table VIII)! We note that the methods described here are only designed to reveal the Class I seven β-strand family of methyltransferases. Further work will be needed to analyze the Class II (SET) enzymes and the Class III (membrane-bound) enzymes. From the compilation in Table VIII of the 38 presently identified yeast methyltransferases, 26, or 68%, are of the Class I type.

Based on our results, it appears that we may have reached the limit of what is possible with the SM^2 methodology presented. Doubling the training set had minimal effect on the results. When we included information from the motif Post I, we did increase the number of correct positive identifications but only marginally improved the number of candidate methyltransferases returned above the 5-false positive threshold used in this study. It is clear that SM^2 may weakly score some methyltransferases (false negatives) because the motifs are divergent or because the spacing between them is different from the canonical spacing.

So how can these results be improved further? The next logical step would be the incorporation of countertraining sets using the false positive results to create a feature set that could be recognized and used to downgrade ORFs that had similar features. For example, many of the false positives either fit into a class of enzymes that could be identified (*e.g.* dehydrogenases or nucleotide-binding proteins) or were highly homologous and could be eliminated on that basis (*e.g.* the HXT proteins). Another avenue we are currently exploring is the use of motif-based profile HMMs that would automate functional assignments and provide more stringent statistical criteria for distinguishing true *versus* false positives.[3] Despite these limitations, we now have a list of unidentified ORFs for which we are highly confident that a majority of the members will ultimately be characterized as methyltransferases.

* This work was supported by National Institutes of Health Grants GM26020 and AG18000 (to S. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains Supplement I.

¶ Special Fellow of the Leukemia and Lymphoma Society.

‖ To whom correspondence should be addressed: UCLA Molecular Biology Institute, 611 Charles E. Young Dr. East, Los Angeles, CA 90095-1570. Tel.: 310-825-8754; Fax: 310-825-1968; E-mail: clarke@mbi.ucla.edu.

REFERENCES

1. Mao, X., Schwer, B., and Shuman, S. (1996) Mutational analysis of the Saccharomyces cerevisiae ABD1 gene: cap methyltransferase activity is essential for cell growth. *Mol. Cell. Biol.* **16,** 475–480
2. Anderson, J., Phan, L., and Hinnebusch, A. G. (2000) The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 5173–5178
3. Lafontaine, D., Delcour, J., Glasser, A. L., Desgres, J., and Vandenhaute, J. (1994) The DIM1 gene responsible for the conserved m6(2)Am6(2)A dimethylation in the 3′-terminal loop of 18 S rRNA is essential in yeast. *J. Mol. Biol.* **241,** 492–497
4. Cheng, X., and Blumenthal, R. M. (eds) (1999) *S-Adenosylmethionine-Dependent Methyltransferases: Structures and Functions*, World Scientific Publishing Co., River Edge, NJ
5. Cheng, X., and Roberts, R. J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.* **29,** 3784–3795
6. Yeates, T. O. (2002) Structures of SET domain proteins: protein lysine methyltransferases make their mark. *Cell* **111,** 5–7
7. Romano, J. D., and Michaelis, S. (2001) Topological and mutational analysis of Saccharomyces cerevisiae Ste14p, founding member of the isoprenylcysteine carboxyl methyltransferase family. *Mol. Biol. Cell* **12,** 1957–1971
8. Ingrosso, D., Fowler, A. V., Bleibaum, J., and Clarke, S. (1989) Sequence of the D-aspartyl/L-isoaspartyl protein methyltransferase from human erythrocytes. Common sequence motifs for protein, DNA, RNA, and small molecule S-adenosylmethionine-dependent methyltransferases. *J. Biol. Chem.* **264,** 20131–20139
9. Kagan, R. M., and Clarke, S. (1994) Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes. *Arch. Biochem. Biophys.* **310,** 417–427
10. Niewmierzycka, A., and Clarke, S. (1999) S-Adenosylmethionine-dependent methylation in Saccharomyces cerevisiae. Identification of a novel protein arginine methyltransferase. *J. Biol. Chem.* **274,** 814–824
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410
12. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
13. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* (*Oxf.*) **14,** 755–763
14. Schubot, F. D., Chen, C. J., Rose, J. P., Dailey, T. A., Dailey, H. A., and Wang, B. C. (2001) Crystal structure of the transcription factor sc-mtTFB offers insights into mitochondrial transcription. *Protein Sci.* **10,** 1980–1988
15. Alexandrov, A., Martzen, M. R., and Phizicky, E. M. (2002) Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. *RNA* (*N. Y.*) **8,** 1253–1266
16. Galardi, S., Fatica, A., Bachi, A., Scaloni, A., Presuti, C., and Bozzoni, I. (2002) Purified Box C/D snoRNPs are able to reproduce site-specific 2′-O-methylation of target RNA in vitro. *Mol. Cell. Biol.* **22,** 6663–6668
17. Bailey, T. L., and Elkan, C. (1994) in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36, AAAI Press, Menlo Park, CA
18. Bailey, T. L., and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* (*Oxf.*) **14,** 48–54
19. Lee, J. H., Cook, J. R., Pollack, B. P., Kinzy, T. G., Norris, D., and Pestka, S. (2000) Hsl7p, the yeast homologue of human JBP1, is a protein methyltransferase. *Biochem. Biophys. Res. Commun.* **274,** 105–111
20. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29,** 2994–3005
21. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280
22. Heurgue-Hamard, V., Champ, S., Engstrom, A., Ehrenberg, M., and Buckingham, R. H. (2002) The hemK gene in Escherichia coli encodes the N(5)-glutamine methyltransferase that modifies peptide release factors. *EMBO J.* **21,** 769–778
23. Nakahigashi, K., Kubo, N., Narita, S., Shimaoka, T., Goto, S., Oshima, T., Mori, H., Maeda, M., Wada, C., and Inokuchi, H. (2002) HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 1473–1478
24. Wilson, G. G. (1992) Amino acid sequence arrangements of DNA-methyltransferases. *Methods Enzymol.* **216,** 259–279
25. Anderson, R. M., Bitterman, K. J., Wood, J. G., Medvedik, O., and Sinclair, D. A. (2003) Nicotinamide and PNC1 govern lifespan extension by calorie restriction in Saccharomyces cerevisiae. *Nature* **423,** 181–185