Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 2818–2823, March 1998
Biochemistry

Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons

RAJEEV AURORA AND GEORGE D. ROSE*

Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205

Communicated by Russell F. Doolittle, University of California, San Diego, La Jolla, CA, December 23, 1997 (received for review October 2, 1997)

ABSTRACT **We have developed a simple procedure to identify protein homologs in genomic databases. The program, called ORF, is based on comparisons of predicted secondary structure. Protein structure is far better conserved than amino acid sequence, and structure-based methods have been effective in exploiting this fact to find homologs, even among proteins with scant sequence identity. ORF is a secondary structure-based method that operates solely on predictions from sequence and requires no experimentally determined information about the structure. The approach is illustrated by an example: Thymidylate synthase, a highly conserved enzyme essential to thymidine biosynthesis in both prokaryotes and eukaryotes, is thought to be used by *Archaea*, but a corresponding gene has yet to be identified. Here, a candidate thymidylate synthase is identified as a previously unassigned open reading frame from the genome of *Methanococcus jannaschii, viz.,* MJ0757. Using primary structure information alone, the optimally aligned sequence identity between MJ0757 and *Escherichia coli* thymidylate synthase is 7%, well below the threshold of sensitivity for detection by sequence-based methods.**

At least 12 genomes now have been sequenced from diverse organisms, with many additions anticipated in coming weeks. How can this wealth of information best be used to address fundamental questions in biology? In particular, how can related protein domains be identified among organisms that diverged during the Cambrian explosion or earlier (1)? The mechanism of protein evolution gives rise to homologous sequences, with attendant redundancy. Computational biologists have exploited this fact in developing powerful recognition tools. Among these, sequence-based methods (2) to recognize homologs are well developed, but sensitivity falters as sequence similarity sinks into the "twilight zone," a threshold near 30% sequence identity (3). Sensitivity can be extended by using information from multiple aligned sequence families (4, 5), local multiple alignment of blocks (6–9), and structure-based fold recognition such as threading (ref. 10 and references therein) and profiles (11).

Here we present a procedure for homolog recognition based on secondary structure prediction. The method is implemented in a computer program called ORF, an acronym for *O*stensible *R*ecognition of *F*olds. Unlike many other fold recognition approaches, ORF requires no three-dimensional template. In brief, ORF operates solely on sequence information to predict the secondary structure of both an unknown protein and all entries in a database of interest and then uses this information in a query-against-all alignment to select likely candidates. The strategy is based on a simple idea: although sequence space is vast, the number of conceivable protein folds is small, of order 5,000 or fewer (12–15). Typically, such folds can be parsed into a linear sequence of repetitive secondary structure elements interconnected by intervening nonrepetitive regions (i.e., helices, β-strands, and everything else), whose orientation in three-dimensions establishes the fold. The order and size of these elements are expected to be similar in homologous proteins. The converse proposition—*viz.,* elements of similar order and size imply similar proteins—is a likely conjecture (16–18), and we adopt it here.

To demonstrate the approach, we applied ORF to a challenging problem, the identification of a thymidylate synthase (TS) from the archeon *Methanococcus jannaschii*. TS, an ancient and highly conserved enzyme (19), is essential in thymidine biosynthesis. Other enzymes in this pathway appear to have been identified in *M. jannaschii* [e.g., see The Institute for Genomic Research (TIGR) web site: http://www.tigr.org], and *Archaea* are believed to use a TS (20), but an authentic gene for the enzyme has yet to be documented in the literature.

By using ORF, we have identified MJ0757 as a likely TS in *M. jannaschii*. The sequence identity between MJ0757 and TS in *Escherichia coli* is 11% when secondary structure is used to guide alignment (and only 7% from sequence alone). Once identified, a candidate of interest can be validated by methods that are independent of the ORF search procedure. Accordingly, supporting evidence is presented, and a three-dimensional model is developed. Our hypothesis that MJ0757 is an authentic TS in *M. jannaschii* now awaits the attention of experimentalists, who alone can assess its validity.

METHODS

In ORF, the secondary structure of a query protein is compared with that of all test proteins in a database of interest. An a-b-c classification of secondary structure is effected, where "a" is α-helix, "b" is β-strand, and "c" is coil (i.e., all else). Then, optimal pairwise secondary structure alignment of query and test proteins is performed, using dynamic programming (21). It remains an open question whether knowledge of the order and size of secondary structure elements is sufficient to identify a three-dimensional fold uniquely. Assertions to the contrary notwithstanding (22), helix capping studies, which reveal a link between secondary and supersecondary structure (23), are suggestive.

In systematic validation tests to be published elsewhere (unpublished work), ORF was applied initially to a set of diverse folds, using secondary structure assignments extracted from known three-dimensional structures (i.e., observed vs. observed matches). Comparisons among proteins with similar architecture resulted in high scores, with few false-positives, and comparisons between proteins with differing architecture

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: TS, thymidylate synthase; PDB, Protein Data Bank.

*To whom reprint requests should be addressed. e-mail: rose@grserv.med.jhmi.edu.

exhibited low scores, with few false-negatives. In greater detail, the entire Protein Data Bank (24), including 22 different representative folds, was reduced to its corresponding a-b-c sequence by identifying secondary structure from coordinates and then rewriting the sequence in an a-b-c alphabet. Then, each representative fold was used, in turn, as a template to search the full set. False-positives are rare above a search score of +60 in a scale ranging from −100 to +100 (using the scoring matrix given below), and all are eliminated upon inclusion of a length filter that excludes candidates differing by more than ±20%. For example, the PDB includes a large number of two-helix fragments that align well against a subsequence in globins, but these false-positives are mismatched conspicuously in overall size. At this level of stringency, recovery of true positives ranges between 30 and 60% for the 22 representative folds, when the reference set of all like folds is taken to be those identified by VAST (10). Almost all false-negatives can be attributed either to incomplete structures (e.g., only α-carbon coordinates or regions of missing density) or domain insertions that cause frame-shifts.

Next, tests were redone with predicted secondary structure assignments (i.e., predicted vs. predicted matches) by using the GOR prediction method (25). However, the approach does not depend critically on the choice of GOR, and we confirmed that several other currently available alternatives would have sufficed. Following the previous procedure for observed vs. observed matches, the PDB, including all 22 representative folds, was reduced to its corresponding a-b-c sequence, but in this instance secondary structure assignments were obtained from predictions based solely on the sequence; coordinates were ignored. Results were similar to those realized previously in observed vs. observed matches. In particular, false-positives are rare above a score of 60, and nearly all are eliminated by a size filter. Recovery of true-positives for the 22 representative folds ranges between 80 and 125% relative to observed vs. observed. Using this approach, many structural homologs with sequence identity as low as 5–10% were detected, regardless of the fold.

It is important to emphasize that the issue of prediction accuracy is not germane when assessing similarity between predicted structures. On first consideration, this crucial point may seem surprising. However, prediction accuracy is a measure of predicted vs. observed matches, whereas homolog identification by ORF depends instead on the accuracy of predicted vs. predicted matches. At an extreme, an observed strand that is incorrectly but consistently predicted to be a helix in both a target of interest and its homologs would still be discriminatory.

Similar folds can embody dissimilar function. For this reason, secondary structure comparison is only used as a first-level screen. Structurally related candidates identified by ORF are subjected to further, independent tests, such as identification of known functional "landmarks" (e.g., residues required for binding and/or catalysis). Although such additional validation can be time-consuming, it remains practicable because only a handful of likely candidates survive the initial ORF screen. The putative TS from *M. jannaschii* identified here illustrates both steps in the approach.

Alignment and Scoring. For an N-residue sequence, predicted secondary structure is represented as a string of length N in the three-letter alphabet: a, b, and c. Alignment of two strings, corresponding to query and test sequences, was performed using dynamic programming (21), with the score matrix:

	a	b	c
a	1	−1	0
b	−1	1	0
c	0	0	1

Gaps in a and b were penalized by 2 points; those in c by 1 point. Using these values, scores were sorted and alignments were evaluated. The percentage of sequence identity was calculated from the optimally aligned secondary structures.

Search for TS in the *M. jannaschii* Genome. The database of *M. jannaschii* ORFs was obtained from The Institute for Genomic Research (TIGR) web site: http://www.tigr.org. Searches were performed by using the TS sequence of both yeast and *E. coli*. MJ0757 scored high in each search, ranking fourth and fifth, respectively (81 for yeast and 63 for *E. coli*.) The number of likely candidates is a steep function of aligned sequence identity. For example, in yeast, the distribution of candidates by percentile includes one in the 90th percentile, four in the 80th (including MJ0757), none in the 70th, 39 in the 60th, 132 in the 55th, and so on. Further examination of the top candidates eliminated all but three, based on size. Of these, only MJ0757 was found to have a residue corresponding to the catalytic cysteine.

Model Building. The program LOOK (Molecular Applications Group, Palo Alto, CA) was used to build a three-dimensional model of MJ0757 from the x-ray-elucidated *E. coli* structure (ref. 26; PDB file 2TSC, chain A, 24). Structural alignment based on predicted secondary structure and multiple sequence alignment from CLUSTAL W (27) were used to guide model building. The extent to which hydrophobes are sequestered and polars are exposed was assessed from this initial model. The model then was refined by using 100 cycles of minimization in LOOK followed by further minimization and simulated annealing by using X-PLOR (28). Coordinates for the final model are available on our web site (http://cherubino.med.jhmi.edu).

RESULTS

M. jannaschii is an obligate anaerobe isolated from a deep-sea vent (29). As the name implies, these archaebacteria are methanogenic, with a normal growth temperature of ≈85°C, and their proteins are thermostable. The *M. jannaschii* genome has been sequenced by Bult *et al.* (30). To our knowledge, previous sequence-based searches of all open reading frames failed to identify a TS.

Do *Archaea* Have a TS? TS, a methyl-transferase, converts dUMP to TMP. TMP is essential in DNA synthesis, and therefore TS is believed to be present in all wild-type cells. However, a TS has yet to be documented in archaebacteria. In both eubacteria and eukaryotes, the methyl group is transferred from tetrahydrofolate to dUMP by TS, generating TMP and dihydrofolate. The folate cofactor is not present in most members of the *Archaea* domain, prompting the existence of a normal TS to be questioned (31). However, methanogens are known to use methanopterin, a modified folate (31). Studies using $^{13}CO_2$ in a number of methanogens find that the pyrimidine biosynthesis pathway inferred from labeling patterns resembles corresponding pathways in eubacteria and eukaryotes (32). Recently, Nyce and White (33) reported the presence of a TS activity in cell lysates of the archaebacteria *M. thermophila* and *Sulfolobus solfataricus*. Finally, homologs of all other enzymes required for pyrimidine biosynthesis appear to have been identified in the *M. jannaschii* genome, with the sole exception of TS.

An ongoing effort to purify TS from *Archaea* is documented in the literature. TS activity in *Methanobacterium thermoautotrophicum* was identified previously by tritium exchange (34), and an N-terminal, 30-residue fragment of the protein was obtained. However, conversion of dUMP to TMP could not be demonstrated in purified fractions. Later, Vaupel *et al.* (35), analyzing a clone of N^5,N^{10}-methenyltetrahydromethanopterin cyclohydrolase, encoded by the gene *mch* in *M. thermoautotrophicum*, identified an upstream open reading frame with an N-terminal sequence identical to the one obtained by Krone *et*

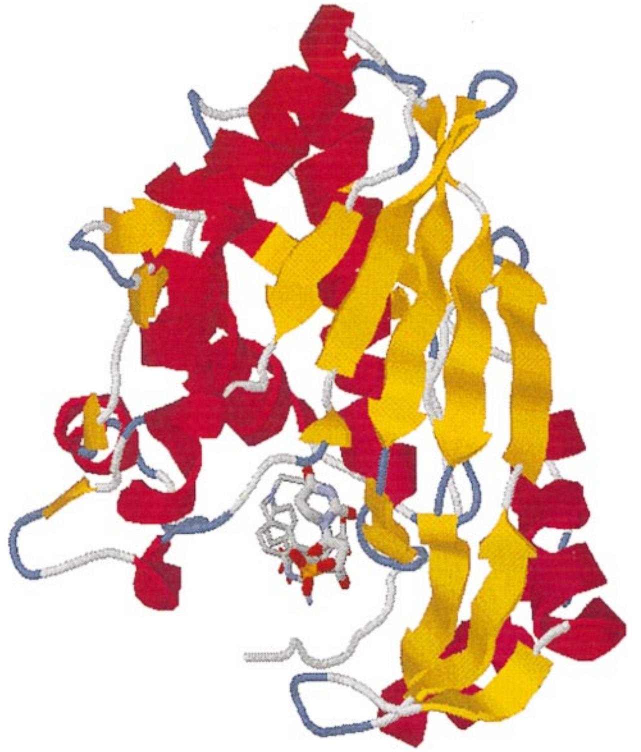


FIG. 1. Cartoon of *E. coli* TS. A ribbon diagram of *E. coli* TS, with helices in red, strands in yellow, turns in blue, and coil in white (PDB file 2TSC, chain A, 26). The active site cleft, on the left, is shown with bound dUMP (thick wireframe) and anti-folate CB3717 (thin wireframe).

al. (34). This finding was poignant because, in both eubacteria and eukaryotes, TS is a tetrahydrofolate-dependent enzyme and genes for TS and dihydrofolate reductase are polycistronic. Vaupel *et al.* (35) tentatively identified this upstream reading frame as *tysY*, the gene for TS. The corresponding *tysY* gene in *M. jannaschii* has been assigned as MJ0511 (30), but the evidence is ambiguous. The *mch* gene appears to be transcribed monocistronically (35). Furthermore, MJ0511 does not score well as a TS, either by our search criteria or by multiple sequence alignment. We conjecture that MJ0511 is likely to be a homolog of deoxyuridylate-hydroxymethyltransferase, a related enzyme found in bacteriophage SPO1 (36). Although *Archaea* are not known to incorporate hydroxymethyluracil in their DNA, the MJ0511 protein may be implicated in tRNA methylation.

Together, the preceding considerations indicate that a TS in *M. jannaschii* will have scant sequence similarity to known homologs, if indeed the enzyme is present at all. Thus, the situation provides an inviting test case for ORF, which, in turn, identified MJ0757 as a likely TS candidate.

Evidence that MJ0757 Is a TS. TS structures from both *Lactobacillus casei* and *E. coli* have been solved by x-ray crystallography (26, 37). Although the two enzymes differ in size (318 residues in *L. casei*, 262 residues in *E. coli*), the two structures have nearly identical architecture. The *E. coli* TS structure is shown as a cartoon in Fig. 1, and its sequence and observed secondary structure are aligned against MJ0757 (260 residues in *M. jannaschii*) in Fig. 2. The secondary structure of MJ0757 predicted by both GOR and PHD (38) is included in Fig. 2.

Evidence that the functionally important sites in TS are found in MJ0757 is described below and summarized in Table 1. In each case, the identification of such residues depends solely on optimal alignment of predicted secondary structure; no further restraint has been imposed. We adopted the convention that TS sequence numbers are keyed to the *E. coli* enzyme, with alternative *L. casei* numbering given in square brackets.

Cys146 [198] is the signature catalytic residue in TS. The thiol acts as a nucleophile that attacks C-6 of dUMP and activates the C-5 carbon for condensation with cofactor (19). After secondary structure alignment, Cys146 [198] coincides

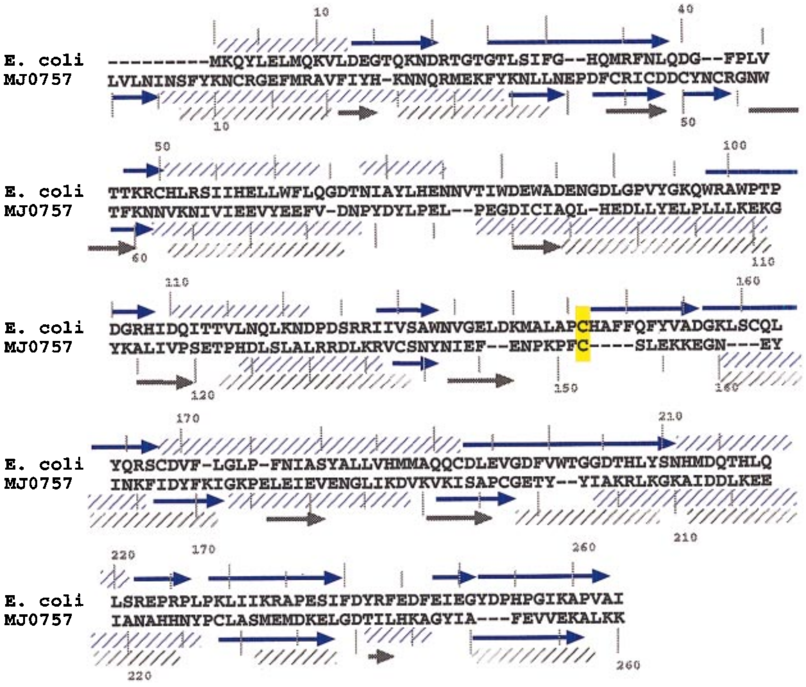

10
40
E. coli ---------MKQYLELMQKVLDEGTQKNDRTGTGTLSIFG--HQMRFNLQDG--FPLV
MJ0757 LVLNINSFYKNCRGEFMRAVFIYH-KNNQRMEKFYKNLLNEPDFCRICDDCYNCRGNW
10
50
50
100
E. coli TTKRCHLRSIIHELLWFLQGDTNIAYLHENNVTIWDEWADENGDLGPVYGKQWRAWPTP
MJ0757 TFKNNVKNIVIEEVYEEFV-DNPYDYLPEL--PEGDICIAQL-HEDLLYELPLLLKEKG
60
110
110
160
E. coli DGRHIDQITTVLNQLKNDPDSRRIIVSAWNVGELDKMALAPCHAFFQFYVADGKLSCQL
MJ0757 YKALIVPSETPHDLSLALRRDLKRVCSNYNIEF--ENPKPFC----SLEKKEGN---EY
120
150
160
170
210
E. coli YQRSCDVF-LGLP-FNIASYALLVHMMAQQCDLEVGDFVWTGGDTHLYSNHMDQTHLQ
MJ0757 INKFIDYFKIGKPELEIEVENGLIKDVKVKISAPCGETY--YIAKRLKGKAIDDLKEE
210
220
170
260
E. coli LSREPRPLPKLIIKRAPESIFDYRFEDFEIEGYDPHPGIKAPVAI
MJ0757 IANAHHNYPCLASMEMDKELGDTILHKAGYIA---FEVVEKALKK
220
260

FIG. 2. Secondary structure alignment of *E. coli* TS and MJ0757. Sequence alignment (in single-letter code) generated by ORF, based on predicted secondary structure. Observed secondary structure (blue), from x-ray coordinates used in Fig. 1, is shown above the *E. coli* sequence, with α-helices indicated by hatched boxes and β-strands by arrows. Secondary structure predicted by both GOR (25) (blue) and PHD (gray) (38) is shown below the MJ0757 sequence. The position of the catalytic cysteine is highlighted in yellow. Residues are numbered for convenience.

Table 1. Equivalent residues in TS from *L. casei, E. coli*, and *M. jannaschii* identified by structural alignment

Role	*L. casei**	*E. coli*†	MJ0757‡
Nucleophile	Cys198	Cys146	Cys152
Phosphate binding	Asp221	Asp169	Asp168
	Arg23	Arg21	Arg29
	Arg218	Arg166	Lys165
Ribose binding	Tyr261	Tyr209	Tyr202?
PABA ring binding	Ile81	Ile79	Leu86
	Leu224	Leu172	Ile172
	Phe228	Phe171	Phe170
	Val314	Val262	Leu258

*From the x-ray structure of *L. casei* TS (37).

†From the x-ray structure of *E. coli* TS (26).

‡From predicted secondary structure alignment of *E. coli* TS against MJ0757, as shown in Fig. 2.

precisely with Cys152 in MJ0757. Among other residues important for dUMP binding and catalysis are two invariant arginines that correspond to an Arg and a Lys in MJ0757. Similarly, residues that bind the PABA ring of folate/methanopterin also are conserved. On the other hand, Tyr209 [261], another invariant residue, fails to align with a tyrosine in the MJ0757 sequence, although two tyrosines (*viz.,* 201, 202) are situated nearby.

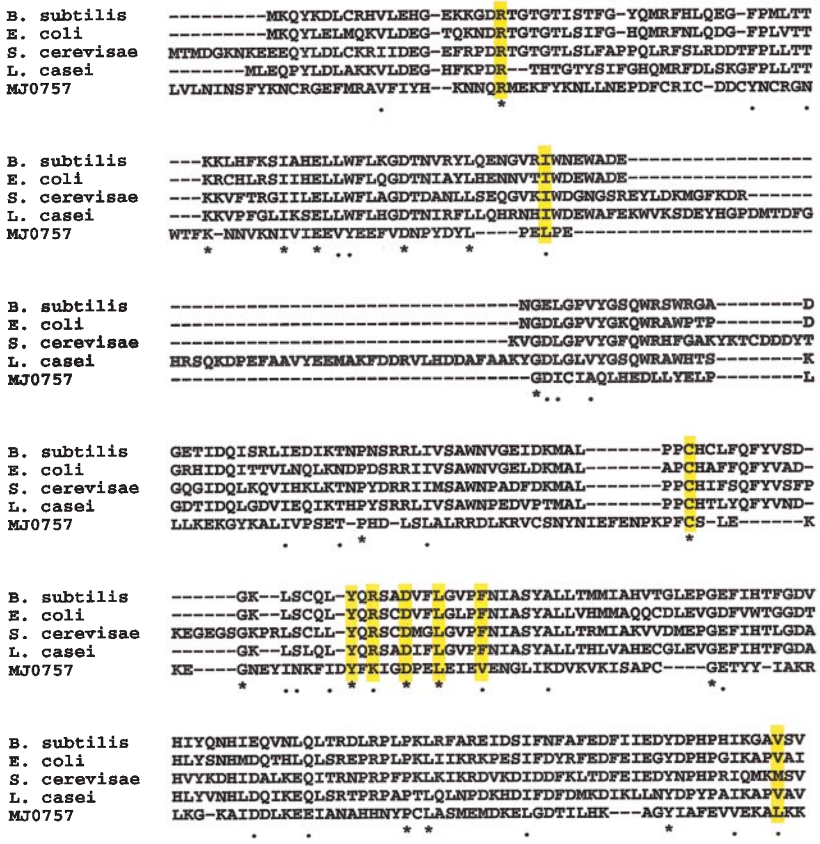
Residues that bind the pterin ring of the folate cofactor are not conserved in MJ0757. We suspect that this fact can be ascribed to differences between folate and methanopterin. In this regard, it is noteworthy that hydrogen bonds between the enzyme and the pterin ring are provided by backbone atoms, and, thus, they are not residue-specific. However, the two tryptophans that pack against the ring are conspicuously absent in the *M. jannaschii* candidate.

Multiple sequence alignment was performed, using TS sequences from *Bacillus subtilis*, *Saccharomyces cerevisiae*, and *L. casei* together with MJ0757. Results from the program CLUSTAL W (27) with default parameters are shown in Fig. 3. Again, residues essential to catalysis coincide, despite the fact that overall sequence identity is low. The best alignment to MJ0757 is obtained with yeast TS, where sequence identity is 16.8%.

As a further independent check, the MJ0757 backbone was threaded onto *E. coli* TS (see *Methods*). The solvent accessibility of polar and hydrophobic residues was assessed in the model and appears to be plausible (Fig. 4), and the binding cavity was preserved upon addition of sidechains to the backbone model. It has been observed that proteins from thermophiles are enhanced in internal salt bridges (refs. 39 and 40 and reference therein) as well as larger side chains that facilitate tighter packing of the core and thereby promote stabilization at higher temperatures. This trend is observed in the TS model of MJ0757, with a higher proportion of β-branched residues in β-sheets than observed in the TS of either *E. coli* or *L. casei*. Also, two internal salt bridges are present although no attempt to restrain these residues was imposed on the model.

Finally, in analogy to two-dimensional gel electrophoresis (where molecular weight vs. pI are used as identifiers), the pI of MJ0757 is calculated to be 5.4, near that of TS from *E. coli* ($\mathrm{pI}_{\mathrm{calc}}$ = 5.5) and *L. casei* ($\mathrm{pI}_{\mathrm{calc}}$ = 5.3). In general, pI values ranging from 4.7 to 5.2 have been reported for TS (41, 42). The combined filters of molecular mass and pI can bracket a protein of interest within a surprisingly narrow range. In the *M. jannaschii* genome, there are only 15 open reading frames with a calculated pI between 5.0 and 5.7 that lie in the size range 29.5–31.0 × 10^3 Da.

Issues of speed are of concern when searching whole genomes. In its current implementation on a workstation of



```
B. subtilis     ---------MKQYKDLCRHVLEHG-EKKGDRTGTGTISTFG-YQMRFHLQEG-FPMLTT
E. coli         ---------MKQYLELMQKVLDEG-TQKNDRTGTGTLSIFG-HQMRFNLQDG-FPLVTT
S. cerevisae    MTMDGKNKEEEQYLDLCKRIIDEG-EFRPDRTGTGTLSLFAPPQLRFSLRDDTFPLLTT
L. casei        -------MLEQPYLDLAKKVLDEG-HFKPDR--THTGTYSIFGHQMRFDLSKGFPLLTT
MJ0757          LVLNINSFYKNCRGEFMRAVFIYH--KNNQRMEKFYKNLLNEPDFCRIC-DDCYNCRGN
                                   .          *                      .    .

B. subtilis     ---KKLHFKSIAHELLWFLKGDTNVRYLQENGVRIWNEWADE-----------------
E. coli         ---KRCHLRSIIHELLWFLQGDTNIAYLHENNVTIWDEWADE-----------------
S. cerevisae    ---KKVFTRGIILELLWFLAGDTDANLLSEQGVKIWDGNGSREYLDKMGFKDR------
L. casei        ---KKVPFGLIKSELLWFLHGDTNIRFLLQHRNHIWDEWAFEKWVKSDEYHGPDMTDFG
MJ0757          WTFK-NNVKNIVIEEVYEEFVDNPYDYL----PELPE----------------------
                   *      *  * ..    *     *      .

B. subtilis     --------------------------------NGELGPVYGSQWRSWRGA--------D
E. coli         --------------------------------NGDLGPVYGKQWRAWPTP--------D
S. cerevisae    -------------------------------KVGDLGPVYGFQWRHFGAKYKTCDDDYT
L. casei        HRSQKDPEFAAVYEEMAKFDDRVLHDDAFAAKYGDLGLVYGSQWRAWHTS--------K
MJ0757          ---------------------------------GDICIAQLHEDLLYELP--------L
                                                 *..  .

B. subtilis     GETIDQISRLIEDIKTNPNSRRLIVSAWNVGEIDKMAL-------PPCHCLFQFYVSD-
E. coli         GRHIDQITTVLNQLKNDPDSRRIIVSAWNVGELDKMAL-------APCHAFFQFYVAD-
S. cerevisae    GQGIDQLKQVIHKLKTNPYDRRIIMSAWNPADFDKMAL-------PPCHIFSQFYVSFP
L. casei        GDTIDQLGDVIEQIKTHPYSRRLIVSAWNPEDVPTMAL-------PPCHTLYQFYVND-
MJ0757          LLKEKGYKALIVPSET-PHD-LSLALRRDLKRVCSNYNIEFENPKPFCS-LE------K
                          .    . *     .                       *

B. subtilis     ------GK--LSCQL-YQRSADVFLGVPFNIASYALLTMMIAHVTGLEPGEFIHTFGDV
E. coli         ------GK--LSCQL-YQRSCDVFLGLPFNIASYALLVHMMAQQCDLEVGDFVWTGGDT
S. cerevisae    KEGEGSGKPRLSCLL-YQRSCDMGLGVPFNIASYALLTRMIAKVVDMEPGEFIHTLGDA
L. casei        ------GK--LSLQL-YQRSADIFLGVPFNIASYALLTHLVAHECGLEVGEFIHTFGDA
MJ0757          KE----GNEYINKFIDYFKIGDPELEIEVENGLIKDVKVKISAPC----GETYY-IAKR
                      *   ..  . * .  *  *   .     .              *.

B. subtilis     HIYQNHIEQVNLQLTRDLRPLPKLRFAREIDSIFNFAFEDFIIEDYDPHPHIKGAVSV
E. coli         HLYSNHMDQTHLQLSREPRPLPKLIIKRKPESIFDYRFEDFEIEGYDPHPGIKAPVAI
S. cerevisae    HVYKDHIDALKEQITRNPRPFPKLKIKRDVKDIDDFKLTDFEIEDYNPHPRIQMKMSV
L. casei        HLYVNHLDQIKEQLSRTPRPAPTLQLNPDKHDIFDFDMKDIKLLNYDPYPAIKAPVAV
MJ0757          LKG-KAIDDLKEEIANAHHNYPCLASMEMDKELGDTILHK---AGYIAFEVVEKALKK
                       .    .        * *        .            *     .   .
```

FIG. 3. Multiple sequence alignment of TS with MJ0757. Optimal alignment of TS sequences *B. subtilis*, *E. coli*, *S. cerevisae*, *L. casei*, and MJ0757 was generated with CLUSTAL W (27), using default parameters. The catalytic Cys and other residues from Table 1 are highlighted in yellow. The 17 absolutely conserved residues are annotated by an asterisk, and other strongly conserved residues are marked by a dot.

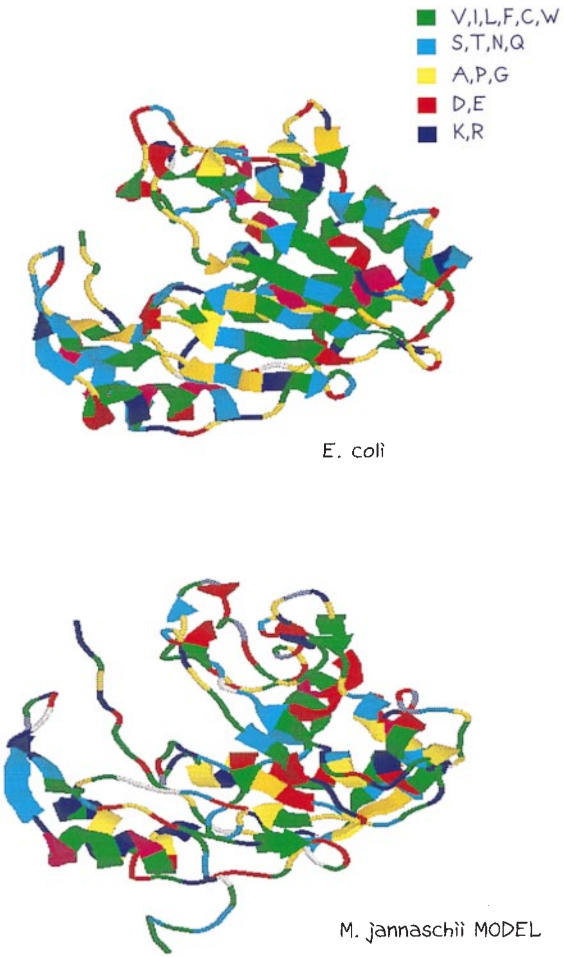


FIG. 4. Three-dimensional model of MJ0757. Using the program LOOK 2.0, the MJ0757 sequence was threaded onto the *E. coli* structure. Side chain placement was refined in LOOK 2.0 and further refined in X-PLOR (28). The model is visualized in RASMOL (53) and color-coded by residue property: hydrophobic (V, I, L, M, F, W, and C), polar (S, T, N, and Q), special backbone (A, P, and G), positively charged (K and R), and negatively charged (D and E). The *E. coli* x-ray structure (*Upper*) and MJ0757 modeled structure (*Lower*) are shown separately for comparison.

medium speed (SGI R4400 Indigo), ORF can search the *M. jannaschii* genome in 7–8 min and GenBank, which contains a quarter-million open reading frames, in 8 h.

DISCUSSION

We have developed a search tool, called ORF, that uses predicted secondary structure to detect protein homologs with scant sequence identity. To illustrate the method, a candidate TS (MJ0757) was identified in an ancient organism, *M. jannaschii*. Once identified, the validity of a candidate can be assessed independently, without further involvement of the search tool. Indeed, when such an assessment is made, many TS residues known to be essential in catalysis and binding are found to be conserved in the identified candidate. The optimally aligned sequence identity between MJ0757 and any TS of known structure is well below the threshold required for detection by traditional sequence-based methods.

Two related structure-based approaches to fold recognition have met with considerable success: profiles and threading. Profile-based methods (see, e.g., refs. 11 and 43) are based on a property matrix (i.e., a *profile*) that is computed for a query sequence and each entry in a library of known structures. Typical properties include secondary structure, solvent accessibility, and residue contact energies. Dynamic programming then is used to identify the optimal match between the profile value of the query sequence and corresponding values for entries from the library. A related approach based on hidden Markov models uses aligned multiple sequence families to develop a statistical profile (44).

In threading, a query sequence is built (i.e., threaded) onto a template of known structure (10, 45). Typically, a library of such templates is tried, and each is evaluated and ranked by using a pseudo-energy potential. Templates with sufficiently favorable scores represent preferred matches. Recent analysis by several groups has raised questions about the validity of threading potentials (46–49). Nevertheless, profiles and threading emerged as the most successful predictive methods in the recent Critical Assessment of Structure Prediction (CASP2) meeting (50). The performance of both methods improves with increased structural similarity between the query sequence and the template (50).

Does Secondary Structure Imply Tertiary Structure? ORF is based on the premise that the secondary structure of a protein determines its tertiary structure, at least in large part. This beguiling premise has cycled through the folding literature for years and has been applied successfully in recent studies (16–18, 43, 51), despite cautionary argument (22). Additional support for this premise comes from our analysis of helix capping (23).

The application of structure to genomics can make a substantial difference in the search strategy. As an example, the evaluation of point mutations is expected to be inherently context-dependent. Replacement of Asp by Glu is a conservative mutation in a β-strand but not in an α-helix (see, e.g., ref. 52). Important to note, such differences are reflected in the GOR (or any equivalent) procedure and therefore in ORF, but they are neglected in the substitution matrices of typical alignment procedures.

A mature science of structure-based genomics would use the predicted three-dimensional structure of the query sequence and a library of known or predicted three-dimensional structures for all relevant genomes. However, predicted secondary structure is a more realistic goal at present, and, as exemplified by TS, it may be sufficient for use in comparative genomics.

We are grateful to Jeremy Berg, who suggested TS as a useful test case. We thank him, Rajgopal Srinivasan, Teresa Przytycka, Venkatesh Murthy, and an anonymous referee for many helpful suggestions. This work was supported by grants from the National Institutes of Health.

1. Doolittle, R. F. (1995) *Annu. Rev. Biochem.* **64,** 287–314.
2. Gribskov, M. & Devereux, J. (1992) in *Sequence Analysis Primer* (Freeman, New York).
3. Doolittle, R. F. (1986) *Of Urfs and Orfs* (Univ. Sci. Books, Mill Valley, CA).
4. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
5. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
6. Smith, H. O., Annau, T. M. & Chandrasegaran, S. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 826–830.
7. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12091–12095.
8. Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997) *Nucleic Acids Res.* **25,** 1665–1677.
9. Henikoff, S. & Henikoff, J. G. (1997) *Protein Sci.* **6,** 698–705.
10. Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996) *Curr. Opin. Struct. Biol.* **6,** 377–385.
11. Luthy, R., Bowie, J. U. & Eisenberg, D. (1992) *Nature (London)* **356,** 83–85.
12. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
13. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994) *Nature (London)* **373,** 631–634.
14. Holm, L. & Sander, C. (1996) *Science* **273,** 595–603.

15. Wang, Z. X. (1996) *Proteins* **26,** 186–191.
16. Russell, R. B., Copley, R. R. & Barton, G. J. (1996) *J. Mol. Biol.* **259,** 349–365.
17. Rost, B., Schneider, R. & Sander, C. (1997) *J. Mol. Biol.* **270,** 471–480.
18. Di Francesco, V., Garnier, J. & Munson, P. J. (1997) *J. Mol. Biol.* **267,** 446–463.
19. Carreras, C. W. & Santi, D. V. (1995) *Annu. Rev. Biochem.* **64,** 721–762.
20. White, R. H. (1997) *J. Bacteriol.* **179,** 3374–3377.
21. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
22. Havel, T. F., Crippen, G. M. & Kuntz, I. D. (1979) *Biopolymers* **18,** 73–81.
23. Aurora, R. & Rose, G. D. (1998) *Prot. Sci.* **7,** 21–38.
24. Bernstein, F. C., Koetzle, T. G., Williams, G., Meyer, E., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
25. Garnier, J. & Robson, B. (1989) in *The GOR Method*, ed. Fasman, G. (Plenum, New York), pp. 417–465.
26. Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. & Stroud, R. M. (1990) *Biochemistry* **29,** 6964–6977.
27. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
28. Brünger, A. T. (1996) in *X-PLOR, Version 3.8. A System for X-Ray Crystallography and NMR* (Yale Univ., New Haven, CT).
29. Jones, W. J., Leigh, J. A., Mayer, F., Woese, C. R. & Wolfe, R. S. (1983) *Arch. Microbiol.* **136,** 254–261.
30. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al.* (1996) *Science* **273,** 1058–1073.
31. White, R. H. (1993) *J. Bacteriol.* **175,** 3661–3663.
32. Choquet, C. G., Richards, J. C., Patel, G. B. & Sprott, G. D. (1994) *Arch. Microbiol.* **161,** 471–480.
33. Nyce, G. W. & White, R. H. (1996) *J. Bacteriol.* **178,** 914–916.
34. Krone, U. E., McFarlan, S. C. & Hogenkamp, H. P. (1994) *Eur. J. Biochem.* **220,** 789–794.
35. Vaupel, M., Dietz, H., Linder, D. & Thauer, R. K. (1996) *Eur. J. Biochem.* **236,** 294–300.
36. Wilhelm, K. & Rüger, W. (1992) *Virology* **189,** 640–646.
37. Finer-Moore, J., Fauman, E. B., Foster, P. G., Perry, K. M., Santi, D. V. & Stroud, R. M. (1993) *J. Mol. Biol.* **232,** 1101–1116.
38. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
39. Vogt, G. & Argos, P. (1997) *Fold Des.* **2,** S40–S46.
40. Jaenicke, R., Schurig, H., Beaucamp, N. & Ostendorp, R. (1996) *Adv. Protein Chem.* **48,** 181–269.
41. Dunlap, R. B., Harding, N. G. & Huennekens, F. M. (1971) *Ann. N Y Acad. Sci.* **186,** 153–165.
42. Haertle, T., Wohlrab, F. & Guschlbauer, W. (1979) *Eur. J. Biochem.* **102,** 223–230.
43. Rice, D. W. & Eisenberg, D. (1997) *J. Mol. Biol.* **267,** 1026–1038.
44. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.* **6,** 361–365.
45. Torda, A. E. (1997) *Curr. Opin. Struct. Biol.* **7,** 200–205.
46. Crippen, G. M. (1996) *Proteins Struct. Funct. Genet.* **26,** 167–171.
47. Thomas, P. D. & Dill, K. A. (1996) *J. Mol. Biol.* **257,** 457–469.
48. Godzik, A. (1995) *Protein Eng.* **8,** 409–416.
49. Godzik, A. (1996) *Protein Sci.* **5,** 1325–1338.
50. Marchler-Bauer, A. & Bryant, S. H. (1997) *Trends Biochem. Sci.* **22,** 236–240.
51. Fischer, D., Tsai, C.-J., Nussinov, R. & Wolfson, H. (1995) *Protein Eng.* **8,** 981–997.
52. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47,** 45–148.
53. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **220,** 374.