

PACRAT: a database and analysis system for archaeal and bacterial intergenic sequence features

William C. Ray* and Charles J. Daniels¹

Children's Research Institute and The Department of Pediatrics, The Ohio State University, 700 Children's Drive, Columbus, OH 43205, USA and ¹The Department of Microbiology, The Ohio State University, 484 West 12th Avenue, Columbus, OH 43210, USA

Received August 14, 2002; Accepted August 22, 2002

ABSTRACT

Analysis of intergenic sequences for purposes such as the investigation of transcriptional signals or the identification of small RNA genes is frequently complicated by traditional biological database structures. Genome data is commonly treated as chromosome-length sequence records, detailed by gene calls demarcating subsequences of the chromosomes. Given this model, the determination of non-called subsequences between any gene and its nearest neighbors requires an exhaustive search of all gene calls associated with the chromosome. Further compounding the issue, the location of intergenic regions for many called genes cannot be resolved unambiguously due to uncertainties in gene boundaries, as well as the presence of other conflicting gene calls. To address these difficulties we have constructed the PACRAT (<http://www.biosci.ohio-state.edu/~pacrat/>) database system. PACRAT preprocesses GenBank genome submissions, evaluates for every gene the character of its relationship to those genes nearest to it, and produces a relationally linked model of the gene ordering for the genome. Using this information, the interface allows the researcher to query gene data as well as intergenic sequence data based on a number of criteria. These include the ability to filter searches based on the status of start and stop positions, or upstream/downstream sequences as conflicting with called genes and automated extension of upstream or downstream searches to find probable operon promoters or terminators. The database is also indexed by KEGG classification, allowing, for example, functionally-related groups of high-quality promoter-containing regions to be easily retrieved as a group.

INTRODUCTION

The PACRAT system (1) is an integrated database, data warehousing and data analysis system. It was designed to simplify the task of acquiring and analyzing functionally correlated sets of sequences for potential biologically relevant patterns.

Rationale

It is well understood that the analysis of functionally related sequences for conserved patterns is complicated by degeneracy in the patterns themselves. This understanding is explicitly codified in traditional genomic databases by the inclusion of uncertainty-quantifying statistics such as E-scores with respect to nearest-neighbor functional annotations, or by the use of qualifying adjectives such as 'putative' or 'potential'. Uncertainties with respect to gene boundaries, however, have not traditionally been codified in databases that treat genomes as chromosome-length sequences with genes indicated as subregions.

This information is however, particularly critical to researchers interested in investigating intergenic regions, or near-gene-start regions for sequences that may be functionally active, such as transcription signals or genes encoding small RNAs. It is especially important in the identification of small RNA genes, where their numbers may equal or exceed those encoding tRNAs (2,3). It is also of relevance to the examination of protein gene sequences themselves, as the presence of miscalled starts in a database may erroneously include non-gene sequence in, or exclude valid sequence from, statistical analyses of gene coding regions.

Further complicating the issue, such a treatment of genome and gene data makes the acquisition of intergenic regions difficult. In this model, there is no indication of whether a gene boundary overlaps another called gene, and while gene entries are organised in order of increasing start coordinate, probable miscalls result in situations where non-neighboring gene entries must be examined to detect possible conflicts. Because of this, retrieval of the intergenic regions associated with any gene requires the examination of not only the

*To whom correspondence should be addressed. Tel: +1 6147222557; Fax: +1 6147223273; Email: ray.29@osu.edu

Table 1. PACRAT classifications of archaeal genomes in GenBank

Intergenic region character	PAC-RAT category	<i>Aeropyrum pernix</i> (6)	<i>Archaeoglobus fulgidus</i> (7)	<i>Halobacterium</i> sp. NRC-1 (8)	<i>Methanococcus jannaschii</i> (9)	<i>Methanobacterium thermoautotrophicum</i> (10)	<i>Methanopyrus kandleri</i> AV19 (11)
Bounded by same-strand genes, large in size	0	607	656	976	686	635	457
Bounded by divergent genes, large in size	1	464	638	900	503	448	408
Same-strand bounds, small in size (Probable operon structure)	2	228	418	261	391	509	361
No IGR—same-strand genes have stop-start overlap	4	52	328	237	87	177	165
Short IGR, divergent genes (Possible miscalled start)	5	20	34	52	14	24	132
No IGR—divergent genes have short start-start overlap (Possible miscalled start)	7	6	6	66	2	8	6
No IGR—same-strand genes have long stop-start overlap (Possible miscalled start or miscalled ORF)	8	188	329	107	133	117	165
No IGR—divergent genes have long start-start overlap (Possible miscalled start or miscalled ORF)	9	327	54	70	12	2	28
Bounded by convergent genes, large in size	1	395	214	492	264	380	262
IGR bounded by convergent ORFS, small in size	3	104	146	425	166	72	68
No IGR, short stop-stop overlap (possible miscalled ORF)	7	93	170	104	72	18	88
No IGR, long stop-stop overlap (possible miscalled ORF)	9	223	202	64	30	12	154

immediately neighboring gene entries, but potentially exhaustive examination of the other gene entries for the chromosome.

Since gene boundaries are unlikely to be exactly determined without experimentally mapping each translated gene sequence, we have developed the PACRAT system to catalog and classify the characteristics of the boundaries of genes from GenBank (4) submissions and to allow researchers interested in these sequences to retrieve genes, or the intergenic sequences related to them, based on the characteristics of their called-boundary relationships to all other nearby genes.

It is also important in such analyses to be able to conveniently retrieve groups of sequences based on proposed functional relationships. Therefore, providing multi-sequence

retrieval and analysis functions through proposed functional mappings such as KEGG (5) classification has been an additional focus of the project.

Database design

The PACRAT system is a relational database built using the MySQL (DataKonsultAB) SQL database system. The system loads GenBank whole-genome submissions in .gbk format, and their related KEGG gene classification tables. At load time, a bi-directionally linked list is built from the individual gene regions indicated in the GenBank file. Linkage is based upon the immediately preceding and following called genes in

Table 1. Continued

<i>Methanosarcina acetivorans</i> str C2A (12)	<i>Methanosarcina mazei</i> Goel (13)	<i>Pyrobaculum aerophilum</i> (14)	<i>Pyrococcus abyssi</i> (15)	<i>Pyrococcus furiosus</i> DSM3638 (16)	<i>Pyrococcus horikoshii</i> (17)	<i>Sulfolobus solfataricus</i> (18)	<i>Sulfolobus tokodaii</i> (19)	<i>Thermoplasma acidophilum</i> (20)	<i>Thermoplasma volcanium</i> (21)
2070	1721	763	471	577	478	913	865	523	521
1424	1052	896	522	594	506	1006	906	566	548
369	312	353	347	431	334	363	313	195	237
203	139	313	218	227	190	237	228	131	144
26	6	74	6	24	16	34	60	4	0
8	0	6	2	4	6	4	4	0	0
208	133	273	192	254	180	395	338	105	98
21	13	17	19	57	140	59	161	6	0
1384	1032	396	132	156	208	558	485	300	238
44	22	128	80	130	82	126	100	158	184
6	4	182	106	154	148	136	118	48	60
45	15	289	231	239	230	283	425	70	68

the case of non-overlapping genes, or upon the gene with the largest exterior extent for overlapping genes. From the bi-directionally linked list, the relationship between each gene, in terms of proximity or potential conflict between it and its nearest neighbors is determined. The bi-directional linkage is stored as relational data in an SQL table, along with codes indicating the gene-call's positional relationship to its neighbors. Also classified is an impact range for the gene, detailing the genome-coordinate extent over which non-explicitly linked data must be examined to retrieve other potentially interesting or related features. This feature is useful not only in allowing the system to retrieve useful information regarding surrounding features when queried for a gene, it also provides the

ability to easily determine the set of genomic features that may be of interest for any particular genomic coordinate. Another table includes annotation data related by gene id. For retrieval efficiency, sequences extracted for the gene, upstream (pre-gene) and downstream (post-gene) intergenic regions and 100 bp sequences preceding and following the called start and stop respectively are also stored and linked by gene id.

THE DATA

The data that are available from the PACRAT system included gene nucleotide sequences and protein sequence translations,

as well as the intergenic regions associated with a gene, 100 bp regions immediately preceding or following a gene, or user-specified variable regions surrounding gene called starts or stops in the range of -99 to +99 around the respective boundary. Also available is a quick overview of the genome region local to a gene showing other sequence features in the immediate area. This display includes notation indicating the identity and positional extents of genes in the locale, as well as other marked sequence features (misc_feature, repeat_region, etc.) in the GenBank file. Through the use of the bi-directional linkage data, the system is capable of chaining backwards through apparent operon structures to determine the upstream boundary for the operon and to return the sequence relative to this point rather than the gene immediately queried, if the user desires. Likewise, it can chain forwards to return operon-related terminator regions.

This data may be searched by gene identifier, description or product name as annotated, or by KEGG functional classification. The retrieval results may be filtered to only include results that do, or do not, display any particular PACRAT classification of start or stop relationship with its neighbors. Sequences are returned in multi-sequence FASTA format, and may, through other facilities of the PACRAT system, be analysed and stored for future use directly in PACRAT.

Due to the particular interest in archaeal transcription as a model of eucaryal Pol II transcription, with the corresponding requirement for intergenic sequence data, the PACRAT system currently focuses on complete archaeal genomes that have been submitted to GenBank. All archaeal genomes available as complete genomes from GenBank as of July 2002 are currently available from the system. Bacterial genomes can be loaded into the system on request, and several are currently available. Genomes are periodically reloaded to take advantage of updated information in the GenBank or KEGG databases.

Characteristics

The archaeal genomes in the database currently display PACRAT intergenic region (IGR) classifications as shown in Table 1. The assumption that significant overlaps between genes, or between genes and regulatory regions do not occur is certainly not universally correct. However, the population of called genes with these apparent characteristics is generally small in the Archaea. This scarcity brings into question the boundary calls for those genes that do show significant overlaps. Of note, several of the genomes listed in Table 1 have been significantly reannotated by NCBI. In one, the data as originally submitted showed greater than half the genes in the genome as having called-boundary conflicts of significant length. These numbers have been significantly reduced in the NCBI reannotation.

Searching

The PACRAT database is searched through a World Wide Web interface available at <http://www.biosci.ohio-state.edu/~pacrat/>. This interface provides complete access to the functionally-related sequence retrieval engine, and the filtering features of the PACRAT database. The PACRAT system also allows the storage and editing of retrieval results, and MEME (20),

MAST (22) and CLUSTALW (23) analyses of the results directly within the online system. Details regarding the use of these facilities are available from the 'About' page linked from the above URL.

RESULTS

Comparison between retrievals filtered for only clearly non-conflicting upstream regions and those made with upstream regions that have been classified as in conflict with other genes, demonstrates that the elimination of clearly conflicted upstream regions eliminates potentially damaging noise from promoter pattern analyses. Figure 1 shows the results of MEME pattern discovery analyses for the probable promoter regions of several different functional classes of *Archaeoglobus fulgidus* genes. These analyses were performed on sequences that showed no upstream region conflicts. The BRE and TATA elements displayed in the patterns are statistically significant components of archaeal promoters (24). Figure 2 shows the results of a similar MEME analysis performed on the naive upstream regions of *A.fulgidus* carbohydrate metabolism genes that significantly overlap other genes. While the information content of this pattern is

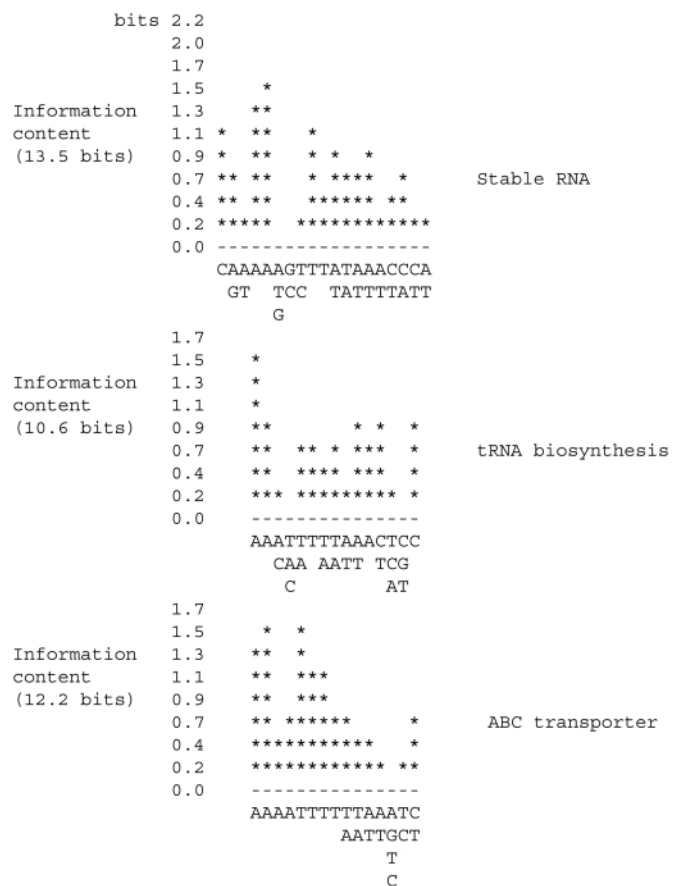


Figure 1. A comparison of the information content and multilevel consensus for promoter-like patterns found in the upstream sequence regions associated with *A.fulgidus* RNA genes, ABC-transporter genes, and tRNA biosynthesis genes. The consensus and information content diagrams are aligned on the leading AA pair followed by the pair of low-information-content positions. These patterns occur consistently near position -25 from the called gene starts.

