



## The PACRAT system: an extensible WWW-based system for correlated sequence retrieval, storage and analysis

William C. Ray\* and Charles J. Daniels

Biophysics Program and Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA

Received on January 31, 1999; revised on August 7, 2000; accepted on August 10, 2000

### ABSTRACT

**Summary:** With PACRAT (Patterns, Analyses, Correlations. Remote Archive Testbed) we present an online database solution to the problem of accessing high-confidence sequences with specific relationships to classes of genes, such as upstream intergenic regions attached to tRNA genes. In addition the software contains a data warehousing and analysis-facilitating suite to streamline the process of analyzing the collected data. An unexpected additional benefit of the system is that it also provides easy access to sequences of lower confidence, and may be of assistance in such things as resolving ORF-call conflicts in genomic annotation projects.

**Availability:** An instance of the PACRAT system, implemented on a set of archaeal genomes, an experimental version of a Yeast genome database, and a random sequence database, is available for use at <http://www.biosci.ohio-state.edu/~pacrat/>. The system is also available for download for installation of local project-oriented sites. See the online PACRAT 'About' page for details.

**Contact:** ray@biosci.ohio-state.edu; daniels.7@osu.edu

**Supplemental information:** Additional information regarding the PACRAT system, including sample datasets is available linked from the 'About' page of the PACRAT website at: <http://www.biosci.ohio-state.edu/~pacrat/>.

### INTRODUCTION

One current standard for definitive public archiving of genomic data is the GenBank database, with access provided through a number of online mechanisms (Benson *et al.*, 1999). World Wide Web (WWW) access is provided through several interfaces directly from the NCBI web pages, as well as through a number of more specific repositories and secondary data providers. Other online access is facilitated through the provision of email interfaces and commandline clients for various platforms. For the researcher interested in detecting conserved

patterns in the sequences between genes however, all of these interfaces suffer from similar difficulties. Not only is the process of retrieving sequences belonging to related families of genes labor intensive, the annotations provided are sometimes inconsistent and do not lend themselves to the detection of such things as conflicting ORF or gene (in GenBank generally 'Coding Sequence' or CDS) calls. Both of these factors can lead to corruption of a discovered pattern, as sequences that are similar but not homologous may adversely affect a derived consensus weight matrix.

For example, an attempt to examine the promoter regions associated with RNA genes from an organism such as *Archaeoglobus fulgidis* via the Entrez genomes database (one NCBI access point to the GenBank data) requires:

- (1) The individual retrieval of 49 GenBank annotated genome sequence regions.
- (2) Searching each of these regions for the coding sequence annotated as belonging to the RNA gene of interest.
- (3) Noting the position and orientation of the RNA gene in each.
- (4) Noting the appropriate terminus of the gene that appears immediately upstream of the RNA gene.
- (5) Extraction of the sequence between this gene and its upstream neighbor's terminus.
- (6) Reverse complementation as necessary of the retrieved sequence.
- (7) Construction of an appropriate sequence database (such as a multiple-sequence FASTA file) for submission to a pattern searching program.

Even if these manipulations are performed successfully and without human error, the researcher still cannot be certain that a sequence retrieved via this procedure is a valid intergenic sequence. The manipulations described do

\*To whom correspondence should be addressed.

not include checking to prevent the accidental extraction of an immediate upstream neighbor in an operon rather than a promoter region. While GenBank sequence regions are annotated in 'generally ascending order', there are instances of CDS region overlaps that can cause what initially appears to be a nearest neighbor to be incorrect<sup>†</sup>.

From early efforts such as the *Methanobacterium thermoautotrophicum* online database (Ray and Daniels, 1997) which annotated FASTA search hits to intergenic regions, but did not allow their direct retrieval, to the more recent *Saccharomyces cerevisiae* database (Chervitz *et al.*, 1999) intergenic sequence BLAST capability, and the G. Church lab upstream region retrieval database (McGuire *et al.*, 1999), researchers interested in intergenic sequence comparison have turned to automation of this procedure with varying success. With the PACRAT sequence database, we extend the intergenic sequence extraction algorithm originally developed for the *M.thermoautotrophicum* database with the notion of 'confidence scores'<sup>‡</sup>, which allow the ranking of intergenic sequences by their probable biological relevance.

In addition to the problem of intergenic sequence retrieval the general record-keeping and housekeeping tasks associated with keeping large collections of sequence data and correlated analysis results must also be addressed.

We present the PACRAT system as an integrated solution to problems in both of these conceptual areas. Initially intended to facilitate the collection of 'high confidence' intergenic sequences from the completed archaeal genomes, it quickly became apparent that streamlining the analysis process, and increasing the reusability of the collected sequences were also important design goals. The focus of the PACRAT system has therefore grown from solely a sequence data retrieval engine to both a retrieval engine and a data warehouse/analysis server. These functions are supported entirely via a web-interface, so as to be accessible from any location. Additionally, they are independent, and either portion of the system can be used separately from the other.

The database portion of the PACRAT system uses GenBank genome description files to build a database of information and sequences related to each CDS in the description. It then categorizes the 'confidence' of CDS starts and stops, and of the intergenic sequences upstream

and downstream of each CDS. The confidence definitions are based on the position of the CDSs neighbors, and the potential biological significance or consequences of the locations of the neighbors. Using this data, and pathway descriptions from the KEGG (Goto and Kanehisa, 2000) pathway annotations, the PACRAT database allows the retrieval of sequence data using the GenBank descriptive fields, or by KEGG annotation. Whether the data is CDS sequence for genes themselves, or for the intergenic sequences before or after these CDSs, it is returned to the user as a single multiple-sequence FASTA format file.

The data warehousing and analysis server portion of the system allows multiple-sequence FASTA format files to be annotated with metadata such as databases searched and user comments, and stored directly on the server. Stored sequences can be submitted to analysis servers, and returned results also archived on the server. These multiple-sequence FASTA files can be uploaded directly from the user's computer, or generated from the output of the PACRAT database system.

The combination of these two functions allows for a variety of uses that range from simplified retrieval and alignment of related proteins, to detection of conserved intergenic sequence patterns, to collaborative sequence analysis and annotation projects.

## IMPLEMENTATION

From the user's perspective, the PACRAT system is a set of data import, export and analysis modules arranged around the centralized data warehouse server. From the point of view of a user interested in pattern analysis and discovery in related sequences, the data warehousing and analysis modules are a convenient accessory to the database engine. A researcher interested in streamlining in-lab sequence analysis and collaboration on the other hand, can use the sequence warehouse and analysis server to provide location independent cross-platform access to a central sequence and software repository.

Because of the general applicability of the PACRAT data warehousing and analysis function, and its relative simplicity compared to the PACRAT sequence database, we will discuss it first.

### PACRAT datasets

An increasing number of online data retrieval systems are providing for some form of remote web-based data and analysis warehousing. GCG corporation's new SeqWeb product (GCG, 1999), the BioNavigator system (eBioinformatics, 1999), and even NCBI's own Blast server system (Madden *et al.*, 1996) are evolving in the direction of facilitating cross-platform data analysis and remote data sharing by allowing the user to store and access data and analysis results remotely via the WWW. These systems provide a new level of convenience for using their

<sup>†</sup> In *Pyrococcus horikoshii*, PHtRNA04 (tRNA-Ala-TGC) would appear to be downstream of PH0218 by examination of the immediately preceding CDS in the GenBank annotation. This RNA gene however is part of the 16S, tRNA-Ala, 23S grouping, obscured in the *Phorikoshii* GenBank annotation by the occurrence of 2 CDSs annotated entirely within the 16S rRNA region, and additionally by an opposite-strand CDS entirely enclosing PHtRNA04.

<sup>‡</sup> It should be noted that the confidence scores currently used are based on an empirical decision regarding the biological relevance of a sequence feature. Though not currently based on statistically determined confidence intervals, extension to include statistical scoring in the determination of confidence scores is a logical next step.

products, but suffer from a number of problems, such as naming uploaded sequence datasets by the original file's filename<sup>§</sup>, or requiring that the user bookmark or keep track of a cryptic name to access their data<sup>¶</sup>.

The PACRAT dataset is fundamental to the PACRAT data warehousing concept. The dataset functionality is provided through a combination of multiple-sequence FASTA files, associated metadata files containing user-supplied or system generated data related to the contents of the sequence files, and a collection of Perl code to provide access and manipulation functions to the user.

User access to datasets is controlled through the standard HTTPD user authentication scheme (ASF, 1999) to provide privacy and security for user data. The addition of an automatic user creation feature allows new visitors to a PACRAT site to instantly create a userid, and immediately begin using the system.

Dataset manipulation functions include the ability to store and delete datasets, concatenate multiple datasets into a single dataset, and to delete sequence items from a dataset.

Since the dataset is stored as a multiple-sequence FASTA file, with associated metadata stored separately, the addition of new dataset manipulation features is simplified—any program that outputs multiple-sequence FASTA format data can be easily adapted to create or manipulate PACRAT datasets.

### PACRAT analyses

Working from the PACRAT datasets are analyses served by the PACRAT system. As the system's immediate research purpose is examination of conserved intergenic sequence patterns, the analyses currently supported are MEME (Bailey and Gribskov, 1998), MAST (Bailey and Gribskov, 1998), and ClustalW (Thompson *et al.*, 1994).

MEME provides pattern discovery through implementation of the Multiple Expectation-maximization for Motif Elicitation method of hidden Markov model generation.

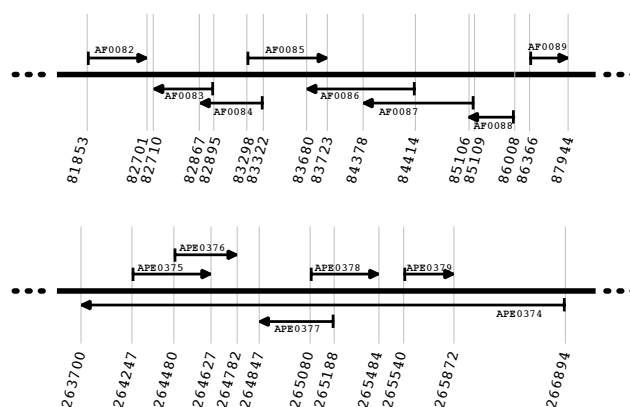
MAST, which takes MEME patterns and multiple-sequence FASTA files as input, uses the consensus weight matrices derived by MEME to search for additional occurrences of a pattern.

ClustalW performs a multiple sequence alignment on FASTA data files.

Again, since PACRAT datasets are simply multiple-sequence FASTA files, addition of other analysis methods that support this format is relatively straightforward.

<sup>§</sup> SeqWeb names uploaded datasets by the original filename, and provides no way for the user to comment, or change this possibly non-informative name.

<sup>¶</sup> NCBI's new format blast server provides persistent analysis results for the user, but requires that the user keep track of an eighteen-digit id number with which to access the data.



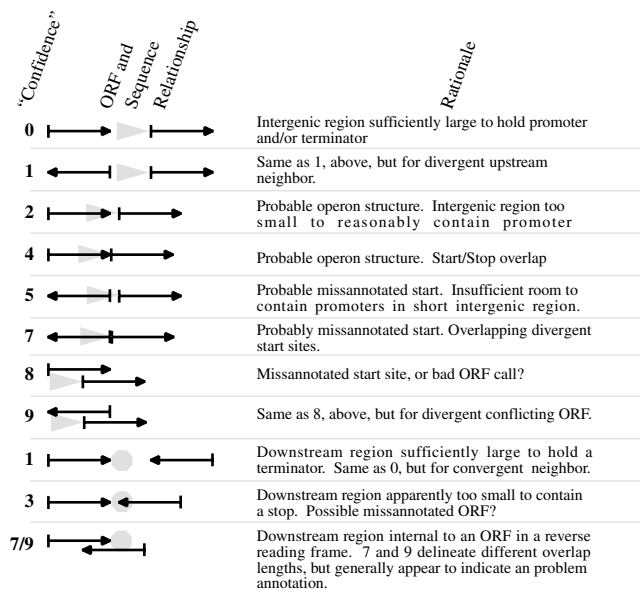
**Fig. 1.** Current NCBI genome annotations contain conflicting data in the form of overlapping CDSs (genes) occurring in the same, or opposite directions. In the *A. fulgidis* region shown, the only potentially promoter-containing region that does not directly conflict with the coding region of another CDS is between the divergent pair of CDSs AF0088 and AF0089. The *A. pernix* example is even less clear. The sequence position markers shown here are not to scale.

### The PACRAT database

The PACRAT database is designed to provide the researcher with the ability to search on, and retrieve data that is not easily accessible through other online database systems. This data includes both sequences that are difficult to retrieve from other databases, sequence 'families' as defined by the KEGG pathways database, as well as certain confidence factors regarding the sequences in question.

The PACRAT database is implemented as a set of SQL tables, running on the MySQL non-commercial SQL server (DataKonsultAB, 1999), and accessed by a collection of integrated Perl scripts and HTML pages.

*Preparing the data.* PACRAT databases are created from GenBank genome description files, and from KEGG pathway databases. The first step in preparing data for the PACRAT system is examination of the data by an extensible rules-based classification system that attempts to determine, given a GenBank genome description file, what CDSs are present, their locations, their nearest upstream and downstream neighbors, and the probable location of sequences such as promoter and terminator containing regions related to each CDS. This classification is made difficult by conflicting CDS definitions in the typical GenBank file, and by the lack of consistency in the use of the GenBank fields between different groups submitting sequences. The CDS maps shown in Figure 1 are from representative regions of the *Archaeoglobus fulgidis* and *Aeropyrum pernix* genomes, and illustrate some of



**Fig. 2.** The confidence score to the left of each representative is assigned to the rightmost, right-facing CDS in the pair, when it is found in the shown relationship to its neighbor. The distances upon which these decisions are based are listed in the main text.

the problems inherent in the annotations. Sequences are then extracted from the GenBank genome description file as appropriate and stored in the SQL database.

**Confidence scores.** Based on each CDSs relationship with its neighbors, the PACRAT classifier assigns to each start and stop, and to each sequence related to the CDS, a set of confidence scores. Confidence scores are in the range [0–10) and are currently assigned integer values from 0 to 9. The integer scores are used as bins into which sequence features are ranked, with a score of zero indicating a feature in which the system has high confidence, and a score of nine indicating the lowest confidence. These confidence scores are then stored in the database to facilitate retrievals with arbitrary cutoffs for the ‘goodness’ or ‘badness’ of the sequences returned. Figure 2 shows the CDS relationship categories on which the confidence scores are based.

The confidence scores are initially based on an automatically calculated score which indicates the possible biological relevance or consequence of the positional boundaries of the sequence. A value is assigned for each of the following:

- The CDS start—whether the position indicated in the database is likely to be biologically correct, or an annotation error. This value is assigned based on the relative positions of the CDSs neighbors as described below.

- The CDS stop—again, whether the position in the database is likely to be correct, or the result of a misannotation. Assigned based on the positions of the CDSs neighbors as described below.
- The CDS itself—whether the CDS is actually likely to be a transcribed sequence, or simply a randomly occurring sequence which triggers a CDS call. The classification system currently ‘believes’ all GenBank annotations are correct. This field is intended to allow researchers to mask false CDS calls out of the database when looking for high-confidence sequences, as might be required for genome annotation type projects.
- The apparent upstream region—whether the sequence directly upstream of this CDS is likely to be the sequence containing the promoter that drives transcription of this CDS. This value is calculated to be identical to the CDS start confidence value.
- The apparent downstream region—whether the sequence directly downstream of this CDS is likely to be the sequence containing the termination sequence for transcription of the CDS. This value is calculated to be identical to the CDS stop confidence value.

The values currently used by the installed instance of the PACRAT database at <http://www.biosci.ohio-state.edu/~pacrat/> for the calculation of the confidence intervals shown in Figure 2 are as follows:

- 30 bp upstream of a start site for that sequence to be considered a type 0 or 1 upstream region.
- 20 bp downstream of a stop site for that sequence to be considered a type 0 or 1 downstream region.
- $\leq 10$  bp overlap in start or stop positions for type 4 or 7 upstream and downstream regions.
- Other confidence scores occur in the ranges between, or beyond these values as shown in Figure 2.

These values are user configurable at database generation time, and are currently optimized to evaluate archaeal gene organization. For development of other organismal databases, these variables may be set to any values that the researcher considers reasonable.

Separate values are included for both the start, and the upstream region (and similarly for the stop/downstream region), even though they are initially assigned the same confidence by the calculation. These are included separately so that a user can intervene and tell the system, for example, that the start site is known with high confidence but that due to conflicting evidence regarding the upstream region, the correctness of the upstream region is still in question.

These confidence values allow, for example, the researcher interested in examining conserved promoter regulatory elements, to limit the sequences returned in a search to only those upstream regions that do not overlap or otherwise conflict with any other CDSs. They also allow chaining backwards, or forwards through the database, searching for the promoter sequences or terminator sequences for an operon.

While the confidence scores are currently used as simple integer bins for ranking sequence features, the system is extensible to allow statistical evaluation of the features and use of this information for ranking as well.

### The PACRAT website

The final facet of the PACRAT system is the web interface to the database, data warehouse, and analysis server. The web interface allows the researcher to retrieve sets of sequences from the database by querying a range of database fields.

Sequences of interest from the returned set can be stored in PACRAT datasets for future use, or datasets can be uploaded directly from the user's computer. Datasets can be combined and edited as necessary, and either directly submitted to the analysis packages available through PACRAT, or retrieved back to the user's local computer as multiple-sequence FASTA format files to be used with analysis packages that PACRAT does not yet support. A more complete description of the use of the database is available from the PACRAT website under the 'About PACRAT' option.

### CONCLUSION

The PACRAT system is evolving at a rapid pace, with new features and more sophisticated retrieval options being added on a regular basis. Even in its nascent state however, PACRAT has already demonstrated its suitability for solving the original problem of selecting intergenic regions appropriate for inclusion in training sets for pattern detection (for sample analyses, see the PACRAT 'About' pages).

The PACRAT system has also proven itself valuable for such purposes as the extraction of high-confidence intergenic sequences without the presence of confounding fragments of gene sequence, for the purpose of cross-

genomic intergenic comparisons (Ray and Daniels, 1999).

We have made available the source for the data warehousing and analysis servers, as well as that for database creation. In addition, we are creating a mailing list for the discussion of PACRAT software improvements, in the hope that open-source development of a genomic database project can prove to be as valuable to biological researchers as such development has been to other community-developed software products. See the online PACRAT 'About' page for current news about the PACRAT system.

### REFERENCES

- ASF, (1999) Apache module mod\_auth. Apache Software Foundation, documented at: [http://www.apache.org/docs/mod/mod\\_auth.html](http://www.apache.org/docs/mod/mod_auth.html).
- Bailey, T.L. and Gribskov, M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
- Benson, D.A. *et al.* (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
- Chervitz, S.A. *et al.* (1999) Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res.*, **27**, 74–78.
- DataKonsultAB, T. (1999) MySQL server and client libraries, available from: <http://www.mysql.org/>.
- eBioinformatics, (1999) BioNavigator Online Bioinformatics Server. Located at: <http://www.bionavigator.com/>.
- GCG, (1999) *SeqWeb*. Genetics Computer Group (GCG), Madison, WI.
- Goto, S. and Kanehisa, M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
- Madden, T.L., Tatusov, R. and Zhang, J. (1996) Applications of network blast server. *Meth. Enzymol.*, **266**, 131–141.
- McGuire, A.M., Hughes, J.D. and Church, G.M. (1999) Discovery of new DNA regulatory motifs in microbial genomes, submitted to Genome Research, Personal Communication.
- Ray, W.C. and Daniels, C.J. (1997) Methanobacterium thermoautotrophicum online database, <http://www.biosci.ohio-state.edu/~genomes/mthermo/>.
- Ray, W.C. and Daniels, C.J. (1999) 3-way cross comparison of Pyrococcus genome intergenic regions, unpublished data.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.