BIOINFORMATICS

Vol. 18 no. 5 2002
Pages 661–665

PGAAS: a prokaryotic genome assembly assistant system

Zhou Yu, Tao Li, Jindong Zhao and Jingchu Luo*

College of Life Sciences and The National Key Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, China

Received on October 4, 2001; revised on December 10, 2001; accepted on December 13, 2001

ABSTRACT

Motivation: In order to accelerate the finishing phase of genome assembly, especially for the whole genome shotgun approach of prokaryotic species, we have developed a software package designated prokaryotic genome assembly assistant system (PGAAS). The approach upon which PGAAS is based is to confirm the order of contigs and fill gaps between contigs through peptide links obtained by searching each contig end with BLASTX against protein databases.

Results: We used the contig dataset of the cyanobacterium *Synechococcus* sp. strain PCC7002 (PCC7002), which was sequenced with six-fold coverage and assembled using the Phrap package. The subject database is the protein database of the cyanobacterium, *Synechocystis* sp. strain PCC6803 (PCC6803). We found more than 100 non-redundant peptide segments which can link at least 2 contigs. We tested one pair of linked contigs by sequencing and obtained satisfactory result. PGAAS provides a graphic user interface to show the bridge peptides and pier contigs. We integrated Primer3 into our package to design PCR primers at the adjacent ends of the pier contigs.

Availability: We tested PGAAS on a Linux (Redhat 6.2) PC machine. It is developed with free software (MySQL, PHP and Apache). The whole package is distributed freely and can be downloaded as UNIX compress file: ftp://ftp.cbi.pku.edu.cn/pub/software/unix/pgaas1.0.tar.gz. The package is being continually updated.

Contact: luojc@pku.edu.cn

INTRODUCTION

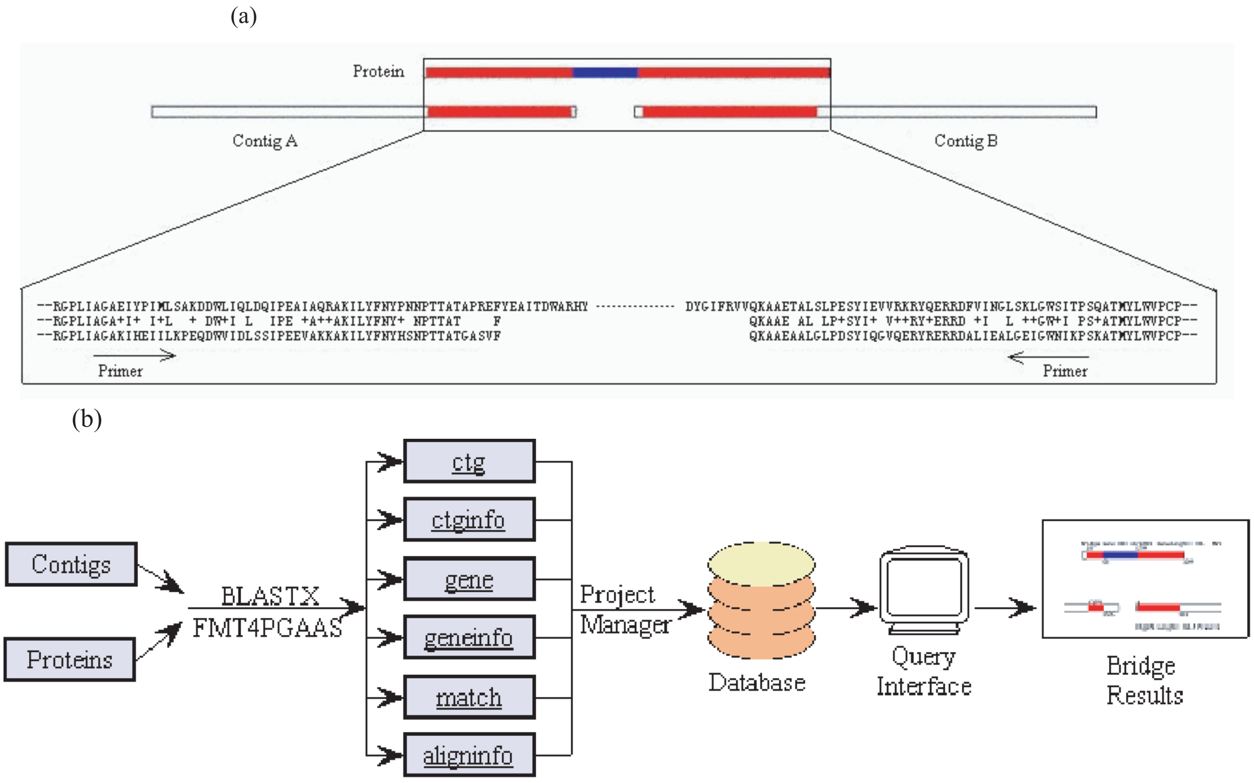
The year 1995 marked the birth of the genome era in biology. During that year, the first complete DNA sequence and annotation of an autonomous living organism, the bacterium *Haemophilus influenzae*, became available (Fleischmann *et al.*, 1995). Since then, more than 60 complete genome sequences have been published, and more than 350 species are being sequenced. The improvement of sequencing technology and the sharp drop of sequencing cost have made large-scale sequencing possible. At the same time, the rapid development of computer technology and bioinformatics has made it possible to assemble very large amounts of sequence data. Driven by these new technologies, shotgun sequencing is now the general approach for large clones and small genomes. This process can be divided into two phases:

1. A number of reads are generated from random subclones and sequencing. They are assembled into contigs after reaching more than six-fold coverage (Lander and Waterman, 1988).
2. The assembly is inspected for validity and for various kinds of data anomalies such as vector sequence contamination and sequence indels. Additional data are collected to close gaps and resolve low quality regions.

The second phase is currently a bottleneck in whole genome sequencing because it is more complex and needs more manual work, and throughput gains depend on reducing the need for human intervention. Several tools for finishing phase have been developed such as the gap4 program (Bonfield *et al.*, 1995) and consed (Gordon *et al.*, 1998). These programs function well as editors to evaluate the base calls and initial assembly, and they use other approaches in the finishing phase rather than using the existing database information. Primer walking is currently a general approach to fulfil this task, which extends the ends of contigs with one primer PCR followed by a sequencing run, but it is time consuming because this PCR approach is single direction with low efficiency (McMurray *et al.*, 1998).

We developed a prokaryotic genome assembly assistant system (PGAAS) package to facilitate contig ordering and gap closure in prokaryotic genome assembly. The basic principle of PGAAS is to find the possible relationships based on the large amount of known protein sequences in public databases. It is known that two species relatives share many genes, therefore the known protein sequences coming from a species relative can be used to provide

*To whom correspondence should be addressed.



```
--RGPLIAGAEIYPIMLSAKDDWLIQLDQIPEAIAQRAKILYFNYPNNPTTATAPREFYEAITDWARHY -------------- DYGIFRVVQKAAETALSLPESYIEVVRKRYQERRDFVINGLSKLGWSITPSQATMYLWVPCP--
--RGPLIAGA+I+ I+L   + DW+I L   IPE +A++AKILYFNY+ NPTTAT    F                              QKAAE AL LP+SYI+ V++RY+ERRD +I   L ++GW+I PS+ATMYLWVPCP--
--RGPLIAGAKIHEIILKPEQDWVIDLSSIPEEVAKKAKILYFNYHSNPTTATGASVF                                  QKAAEAALGLPDSYIQGVQERYRERRDALIEALGEIGWNIKPSKATMYLWVPCP--
```

Fig. 1. (a) The principle of PGAAS. A linkage protein matches two contigs in the BLASTX search, which imitates a bridge standing on two piers. The bottom part shows the sequence alignment between the linkage protein and the two peptides translated from the contigs, which can help in primer design for PCR to close the gap between these two contigs. (b) The flowchart of PGAAS. After running the BLASTX batch job with all contigs against a protein database, the Unix shell script FMT4PGAAS extracts data from the BLASTX output and produces six text files corresponding to the six tables in the PGAAS database schema. These data are loaded to the relational database by the 'Project Manager' and can be used to retrieve bridge information with the graphical query interface.

link information between contigs. If the ends of two contigs match the same protein sequence appropriately, the two contigs are tentatively considered to be adjacent (Figure 1a). The likelihood is higher for prokaryotic species, because there are no introns and few repeats in coding regions. This approach can also be used to other genomes where introns are rare such as organelles.

SYSTEMS AND METHODS

We searched each contig end of the sequences for *Synechococcus* sp. PCC7002 with BLASTX (Altschul *et al.*, 1990) against a peptide database, dumped the results into the PGAAS database and retrieved all the bridge proteins which match more than one contig end, and then designed primers at the joining end of the two pier contigs. With these two primers, the efficiency of PCR will be increased distinctly. With the help of PGAAS, we can find the bridges and piers rapidly, draw images for the matches, and design PCR primers as the last step.

The test contig database comes from an ongoing genome project for *Synechococcus* sp. PCC7002, whose genome size is about 2.7 Mb (Chen and Widger, 1993). A total of 513 contigs were obtained after about six-fold sequencing coverage. The subject protein database we searched was that of a closely related species, *Synechocystis* sp. PCC6803 whose genome has been completely sequenced and well annotated (Kaneko *et al.*, 1996). This database contains 3168 proteins or ORFs.

We tested PGAAS with the default BLASTX parameters except that the 'Expect value' (e) was set to 1×10^{-2} to avoid too many random and low similarity matches, and the number of one-line descriptions in the output file (v) was set to 0.

PGAAS consists of three main parts: a parser to extract information from the BLASTX results, a database to store information extracted by the parser, and an interface to retrieve and display the peptide links and adjacent contigs. The parser FMT4PGAAS is a set of programs including a UNIX shell script and eight Perl scripts. We choose

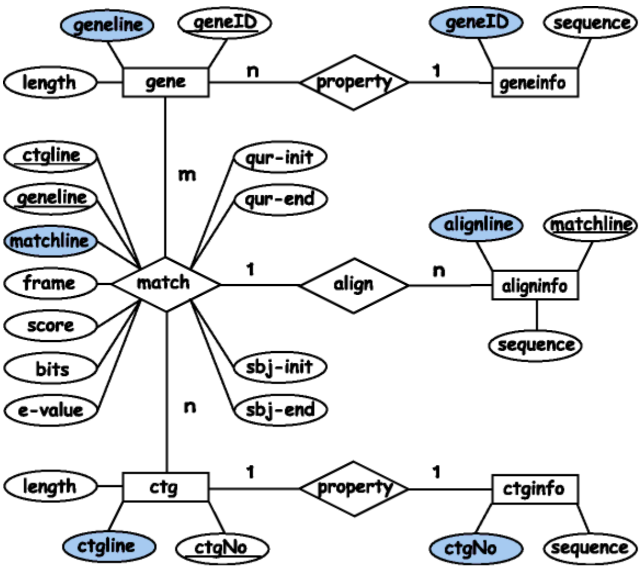


Fig. 2. The entity–relationship (E–R) model of the PGAAS relational database. Different shapes designate different database elements: rectangle box for entities, ellipse for attributes and diamond for relationships. Shaded ellipses denote the primer keys and the underlined attributes are the foreign keys. The numbers '1' and the symbols 'n' and 'm' along the lines denote the relationships between two entities or an entity and a relationship. Six tables were constructed according to the five entities 'gene', 'geneinfo', 'ctg', 'ctginfo' and 'aligninfo' as well as the relationship 'match'. The entity 'gene' contains the proteins and potential ORFs from the BLASTX results and the entity 'geneinfo' contains their sequences. The entity 'ctg' contains the query contigs whose sequences are available from the 'ctginfo' table. Finally, the entity 'aligninfo' contains the BLASTX alignment results line by line. The line numbers in the BLASTX results were used as the unique identifiers of each match.

MySQL as the database management system, PHP4 as the application programming interface and Apache as the http server program. All these tools are freely available and meet the needs of PGAAS.

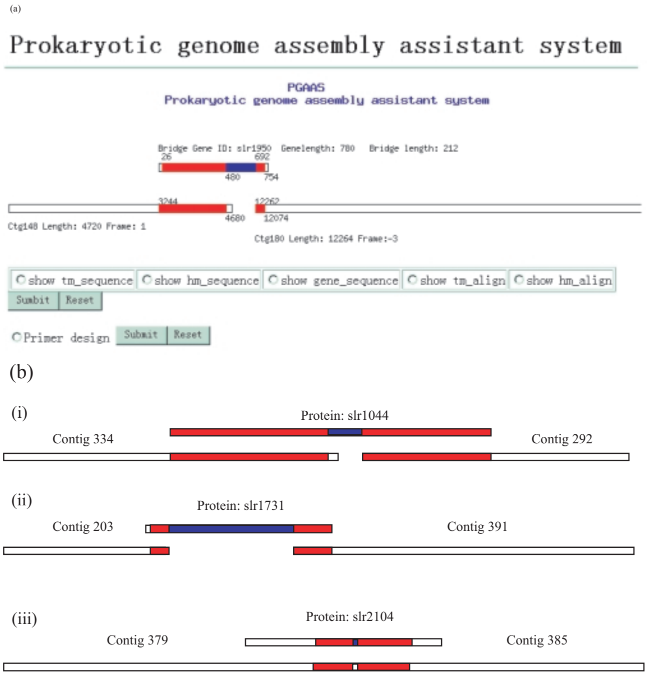
ALGORITHM AND IMPLEMENTATION

Figure 1(b) depicts the flow and relationships among the files and programs described in this paper. The entity–relationship (E–R) model describes the storage of the data and the relationships between the tables (Figure 2). This relationship schema is in the Fourth Normal Form of the relational database theory.

After running BLASTX, users can use FMT4PGAAS to extract the key information from the output file, which is started as a UNIX command with the parameters: the name of the BLASTX output file, the location of both the contig data file and the protein data file in FASTA format. The other operations can be executed through the Web interface. Project manager can be used to create a project for the current job and load the data files generated by the previous steps into database tables. After this step, users can manipulate the queries through the query interface.

The query involves the following steps. First, the matches near the end of contigs are found, and these ends are marked with either a head or tail match stamp (hm or tm) to identify the directionality and contig positions in the bridge-pier map. Then, the bridging peptides matching with two or more contigs with different hm or tm stamps are collected. Finally, the results are filtered in case two pier contigs overlap significantly, or the bridging protein has more than 30 unmatched residues at either terminal. This filtering step is optional but recommended.

The open source 'gd' graphics library (http://www.boutell.com/gd/) was used to design the graphic output of the final result. The original contig and protein sequences and their BLASTX alignment can be displayed on the browser by selecting the corresponding buttons (Figure 3a). We have also integrated the Web version of Primer3 (Rozen and Skaletsky, 1998) into the PGAAS system to design primers matching the pier contigs by sending the results to the CGI interface of Primer3 installed locally.


(a)

Prokaryotic genome assembly assistant system

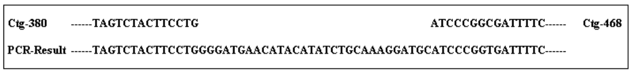
PGAAS
Prokaryotic genome assembly assistant system

Bridge Gene ID: slr1950 Genelength: 780 Bridge length: 212

Ctg148 Length: 4720 Frame: 1

Ctg180 Length: 12264 Frame:-3

Fig. 3. The graphic output of PGAAS. (a) The snapshot of the PGAAS output of peptide link page; the top band is the bridging protein, the two bands below are pier contigs. The grey bands show the matched parts in the BLASTX alignment. (b) Three categories of query results (details are shown in the text).



```
Ctg-380     ------TAGTCTACTTCCTG                                    ATCCCGGCGATTTTC------   Ctg-468
PCR-Result  ------TAGTCTACTTCCTGGGGATGAACATACATATCTGCAAAGGATGCATCCCGGTGATTTTC------
```

Fig. 4. A sequencing experiment in which a pair of adjacent contigs were selected from PGAAS output and primers were designed using Primer3. These two contigs were joined together and the gap was filled successfully using this approach.

RESULTS

The PGAAS outputs can be classified into three categories depending on the length and nature of the gaps (Figure 3b): (i) When an alignment has a high BLASTX score for a substantial portion and the gap is very short, the two contigs can be joined together by the homologous protein with a high level of confidence. (ii) On the contrary, if an alignment has a high BLASTX score for a small portion and the gap is much longer, assigning the sequence according to the homologous bridging protein will be less convincing. (iii) In the third case, the pier contigs can be joined directly if there is no gap between them or they already overlap. To fill the gaps, PCR primers can be designed to match the ends of pier contigs, followed by sequencing of the PCR fragment. The reason that these overlap contigs are not assembled in the initial phase is usually due to the low quality of the terminal sequence(s) or the presence of long vector sequence remaining at the termini. However, PGAAS can not find bridge peptides if the contigs do not have any identifiable homologues in the subject database. These contigs may contain novel genes or domains unique to the new genome being sequenced.

We selected a bridge protein sll0487 from the dataset of *Synechocystis* sp. PCC6803 and the pier contigs (contig 380 and contig 468) and ran PCR using the primers designed by Primer3. The bridge peptide sll0487 is a hypothetical protein which has 59% similarity with the ATP synthase subunits in the cyanobacterium *Synechococcus* 6301. After sequencing the PCR product, the two contigs were joined together in the right order (Figure 4).

DISCUSSION

According to the Poisson distribution equation proposed by Lander and Waterman (1988), the probability that a base is not sequenced in the shotgun sequencing is $P_0 = e^{-m}$, where m is the sequence coverage. Thus, a six-fold coverage yields the probability $P_0 = e^{-6} = 0.00247$, i.e. 0.247%. If L is the genome length and n is the number of contigs, the total gap length is Le^{-m}, and the average gap size is Le^{-m}/n. In the genome of the cyanobacterium *Synechococcus* sp. strain PCC7002, L is 2.7 Mb, thus the unsequenced bases would be less than 6.7 kb. There are 224 contigs less than 2 kb in the total 513 contigs, which are more likely to be misassembled and are accordingly disregarded in this analysis. Therefore, six-fold coverage would leave 289 gaps with an average size of about 25 bp. As a result, the probability that an entire gene falls into the gap regions is very low, and most of the genes should exist in at least one contig. This is demonstrated by the fact that among 3168 proteins or ORFs in the protein dataset of *Synechocystis* sp. PCC6803, 2514 (80%) have matched contigs in the genome of *Synechococcus* sp. PCC7002. The difference represents the genome difference in the two organisms.

Table 1. The parameters and results of testing PGAAS

Sequencing coverage	No. of contigs after assembly	Subject database	No. of tail contigs	No. of head contigs
4	791	PCC6803	103	88
6	513	PCC6803	57	56
6	513	SWISSPROT	36	32

The contigs obtained from the initial assembly of *Synechococcus* sp. PCC7002 was used as the query dataset. The maximum number of unmatched bases at the terminus of each contig is set to 60 bp, the maximum number of unmatched residues at the bridge protein terminus is set to 30aa, the maximal overlap between the left matched part and the right matched part on the bridge is set to 30aa.

The results of PGAAS depend largely on parameter setting, as shown in Table 1. Because some vector segments always remain at the ends of sequencing products, it is necessary to adjust the parameters according to the sequence of the cloning site on the vectors. For example, pUC18 always leaves about 40 bp and BlueScript leaves about 100 bp on the ends, therefore the unmatched bases at the terminus of each contig (Table 1) was set to 60 for pUC18 and 150 for BlueScript to avoid discarding too many positive matched pairs. The maximal overlap between the left matched part and the right matched part on the bridge needs to be adjusted accordingly. After the approximate ranges of the above two parameters are determined, we recommend the most stringent parameters in the first stage to obtain the most confident results, then the restrictions can be relaxed to find more peptide links.

The performance of our approach also depends on the gene distribution of the genome. If the genes are nonuniformly distributed in the genome, contigs in regions with few genes can not be linked through this approach.

PGAAS can also provide some useful information about the bridging proteins in the early phases of sequencing projects. We tested PGAAS when only four-fold sequencing coverage with 791 contigs was available (Table 1). We found more than 50 pier pairs belonging to the first or third category, and one sample pair was tested by experiment with a positive result.

PGAAS may be used in other pre-finished prokaryotic genomes. If the protein sequence database of a closely related species is not available, other genomic protein databases can be used as the subject database. If the SWISSPROT database with multi-gene families is used, the entries within the same gene family should be filtered in the end. We ran a BLASTX search for *Synechococcus* sp. PCC7002 against SWISSPROT (version 39) and ran PGAAS using stringent parameters (Table 1).

In this paper, we describe a software package which can assist the finishing phase of genome sequence assembly. Although the approach of using peptide links to join the contigs has been applied in many genome projects such as the genome sequencing of *Haemophilus influenzae* (Fleischmann *et al.*, 1995), there is no available tool to accomplish the work. We hope that PGAAS can be used to fill this gap. However, PGAAS can not provide information to correct the problem caused by the repeats. It will work well only after the contigs are assembled correctly. The final sequence of the gap of the bridge regions will be obtained by PCR and sequencing.

ACKNOWLEDGEMENTS

We thank Chunwei Wang and Alexander Kamb (Deltagen Proteiomics, Deltagen), and Hua Jing (Harvard Medical School) for their preview and suggestions of the manuscript. This project is supported by the Chinese State Key Basic Research and Development Plan to J. Zhao (G1998010100) and the Beijing Committee of Science and Technology to J. Luo (G9550213500).

REFERENCES

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.

Bonfield,J.K., Smith,K.F. and Staden,R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.

Chen,X. and Widger,W.R. (1993) Physical genome map of the unicellular cyanobacterium *Synechococcus* sp. strain PCC7002. *J. Bacteriol.*, **175**, 5106–5116.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Cougherty,B.A. Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

Kaneko,T., Sato,S., Kotani,H., Tanaka,A. Asamizu,E. *et al.* (1996) Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res.*, **3**, 185–209.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

McMurray,A.A., Sulston,J.E. and Quail,M.A. (1998) Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res.*, **8**, 562–566.

Rozen,S. and Skaletsky,H.J. (1998) Primer3: http://www-genome.wi.mit.edu/genome_software/other/primer3.html.