BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–1

Using MoBIoS' Scalable Genome Join to Find Conserved Primer Pair Candidates Between Two Genomes

Weijia Xu [1], Willard J Briggs[1], Joanna Padolina[2], Wenguo Liu[1], C. Randal Linder[2] and Daniel P. Miranker[*1]*

[1]Department of Computer Sciences, University of Texas at Austin, Austin, TX, 78741, USA
[2]Integrative Biology, University of Texas at Austin, Austin, TX, 78741, USA

ABSTRACT

Motivation: For the purpose of identifying evolutionary reticulation events in flowering plants, we determine a large number of paired, conserved DNA oligomers that may be used as primers to amplify orthologous DNA regions using the polymerase-chain reaction (PCR).

Results: We develop an initial candidate set by comparing the Arabidopsis and rice genomes using MoBIoS (Molecular Biological Information System). MoBIoS is a metric-space database management system targeting life science data. Through the use of metric-space indexing techniques, two genomes can be compared in O(mlog n), where m and n are the lengths of the genomes, versus O(mn) for BLAST based analysis. The filtering of low complexity regions may also be accomplished by directly assessing the uniqueness of the region. We describe mSQL, a SQL extension being developed for MoBIoS that encapsulates the algorithmic details in a common database programming language, shielding end-users from esoteric programming.

Availability: upon request

Contact: {xwj, miranker}@cs.utexas.edu

Keyword: primer pair, genome comparison, SQL, metric space indexing, biological database

1 INTRODUCTION

Biology is difficult, in part, because of the potential interactions among everything that comprises biology. That is to say, every problem addressed by biologists must be considered within its larger context. For *in silico* explorations this usually means developing a new application program for each problem or even problem instance (or at least the ad-hoc scripting and custom parameterization and integration of existing tools). As the importance and challenge of bioinformatics analysis increases so does the need to improve the software-development productivity of biologists by means of more powerful and abstract bioinformatics application development environments. Consider the analogy with the emergence of computers in the commercial world. The 1960's witnessed the original exponential growth of on-line data, and every database application was written one-off. This in-turn led to the emergence of general-purpose database management systems.

MoBIoS (Molecular Biological Information System; pronounced *mobius*) is a next generation database management system targeting life science data (Miranker, et al., 2003). It is currently in development. The MoBIoS architecture embodies metric-space indexing, object-relational models of complex biological data types and an extended SQL query engine capable of mapping concise specifications of bioinformatic problems to good algorithmic solutions. Instead of targeting one particular question, MoBIoS promises to manage large-scale biological data across biological disciplines and enable various specific biological problems to be solved using simple SQL queries (Miranker, et al., 2003). In this paper we present a problem in comparative genomics and detail its solution using MoBIoS.

Phylogenetic trees, commonly used to model evolutionary history, are known to be an over-simplification. In many groups (such as plants) a large proportion of speciation is not tree-like at all (Ellstrand, et al. 1996; Grant, 1971; Rieseberg, 1997). For these organisms speciation occurs on networks as well as trees. In some instances two species can combine their genetic material to produce a completely new species (hybrid speciation). In addition, emerging biological evidence indicates that different chromosomes and different parts of chromosomes in a single species may have vastly different evolutionary histories (Rieseberg, et al.1996; Rieseberg and Linder, 1999). For example, both bacterial and viral sequences are found in the human genome (Venter, et al., 2001). In such an evolutionary scenario, species can contribute genetic material to other species (via horizontal gene transfer and introgression).

To reconstruct an accurate evolutionary history, hybrid speciation and horizontal gene transfer must be identified among species. Reconstruction of hybrid speciation requires multiple, independent, orthologous, nuclear DNA sequences

[*] To whom correspondence should be addressed.

to determine hybrid parentage. However, thus far only a very limited set of regions have been identified for this purpose. One way to obtain these additional regions is to identify a set of universal polymerase chain reaction (PCR) primer pairs across several related species, such that each primer pair amplifies an orthologous region. The sampled material may be sequenced and subjected to genome comparison studies concerning the rates of change in genic and non-genic regions of the amplified DNA in order to determine their suitability for different depths of phylogenetic analysis.

A PCR experiment amplifies the DNA sequence between a pair of oriented primers; one is called the *sense* (forward) primer and the other is called the *antisense* (reverse) primer. For the problem addressed in this paper, we are looking for primer pairs that exist in both the rice and Arabidopsis genomes. A useful primer pair must satisfy a number of properties (Figure-1):

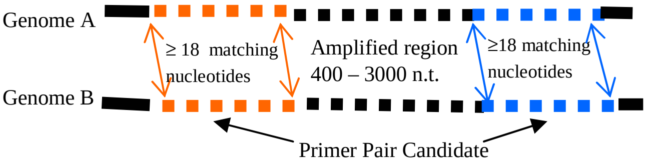

Fig. 1. The specification of a potential primer pair

(1) The lengths of the primers should be between 18 and 25nt to provide a single PCR product that will not amplify multiple genes or exhibit low primer complexity (see below).

(2) The primer sequences must have a minimum of short simple repeats or other low complexity sequences to ensure unique priming.

(3) The primers must not be substantially complementary to each other or themselves to prevent formation of primer dimers and hairpin loops.

(4) The length of the amplified region between the primers in each genome should be between 400 and 3000nt to allow amplification with standard Taq polymerase.

(5) The amplified region must be orthologous in the organisms.

Unfortunately, due to the complicated restrictions on primer pair identification as well as the large amount of sequence data being investigated, neither existing sequence search tools nor primer design tools can provide a simple satisfactory solution without the significant effort of writing a new computer program. Although several programs exist for PCR primer design and selection (Hillier and Green, 1991; Podowski and Sonnhammer, 2001; Li et al., 1997; EMBOSS http://www.uk.embnet.org/Software/EMBOSS/; Wisconsin Package http://www.gc.com/), we are facing a distinctive problem in which no prior knowledge of the targeted genes is available. In fact, the desired amplified regions could differ greatly from one species to another. Our problem also differs from the sequence mapping problem using sequence-tagged sites because we have no knowledge of the primers needed (Schuler, 1997). Therefore, the primer pair candidates must be discovered through genome comparison. It would require substantial effort to develop an optimal solution using existing software tools.

In our previous work we have described the ongoing development of MoBIoS, and our initial focus has been on the investigation of metric models of sequence homology to effectively handle large-scale biological data (Xu and Miranker 2004, Miranker, et al., 2003). Metric-space indexing methods enable O(log n) retrieval time on k-mers based on biological criteria rather than lexical or ordinal keys (Yianilos, 1993; Chavez, et al., 2001; Bozkaya and Ozsoyoglu, 1997, 1999). MoBIoS is implemented in Java and is built on the Mckoi open-source relational database engine (http://www.mckoi.com/). A set of biological SQL extensions (mSQL), analogous to previous spatial extensions, is being developed to abstract these new modeling and algorithm results. It is the addition of mSQL that allows MoBIoS to provide greater flexibility and productivity in solving problems than other computational biology programs specialized for a specific question.

In this paper, we present our solution to the problem of finding conserved regions between two genomes using MoBIoS and algorithms with O(mlogn) time complexity (Section 3). We detail specific results comparing the Arabidopsis and rice genomes and a scalability study (Section 4). We further describe the mSQL language developments and show that complex bioinformatics analysis can be solved by simple SQL programming (Section 5). We conclude that MoBIoS has scalable search performance and that our method is effective at finding biologically interesting regions, which are highly conserved between the rice and Arabidopsis genomes (Section 6).

2 MOBIOS OVERVIEW

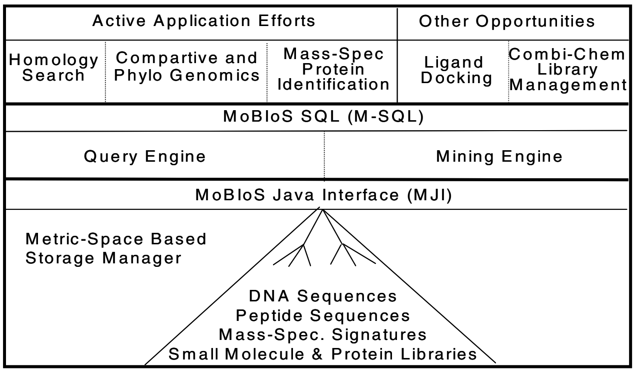
Active Application Efforts
Other Opportunities
Homology Search
Compartive and Phylo Genomics
Mass-Spec Protein Identification
Ligand Docking
Combi-Chem Library Management
MoBIoS SQL (M-SQL)
Query Engine
Mining Engine
MoBIoS Java Interface (MJI)
Metric-Space Based Storage Manager
DNA Sequences
Peptide Sequences
Mass-Spec. Signatures
Small Molecule & Protein Libraries

Fig. 2. Architecture of the MoBIoS Platform.

When relational database management systems are used to manage biological data, important data types are relegated to blob and unstructured text fields. Consequently, mining of the biological data types is accomplished by sequentially dumping the data to utilities outside of the database. Such systems cannot be expected to form optimal solutions or scale in performance as the volume of data increase. For example, in genome comparison a relational database may be used to store the sequences of all of the completely sequenced organisms, but in order to compare genomes A and B, of lengths n and m respectively, the database is limited to returning A and B in their entirety. Commonly BLAST is then used to do the comparison, resulting in an O(mn) algorithm. This is equivalent to a pair of nested loops comparing the complete Cartesian product of the contents of the genomes (Altschul et. al., 1990). When comparing the Arabidopsis and rice genomes, n and m are $1.2x10^8$ and 5.7×10^8 respectively; this yields a very large computation.

MoBIoS solves these problems by incorporating the semantics of biology into its data types, as well as adapting metric-space indexing structures to provide fast pattern matching with sequence substrings. Figure-2 illustrates the architecture of the MoBIoS platform. Metric-space indexes rely only on an ability to measure the similarity of two data objects relative to each other. The single requirement is that similarity be measured using a metric-distance function (a *metric*). No absolute reference coordinates need be used. Thus, the methods can be applied when there is no geometric interpretation of the data, i.e. when the data has no meaning in Euclidean space. Whereas relational database indexes leverage the ordinal character of business data, metric-space indexes may directly integrate biological criteria (Xu and Miranker 2004, Giladi et al., 2002).

Definition 1 Metric Space A metric space is a set of objects with a binary distance function d, satisfying the following conditions for every three objects x, y, and z (Chavez, 2001):

i. d(x,y) ≥ 0 and d(x,y) = 0 iff x = y; (Positivity)

ii. d(x,y) = d(y,x); (Symmetry)

iii. d(x,z) + d(y,z) ≥ d(x,y). (Triangle Inequality)

Metric-space indexing exploits the intrinsic clustering of a data set. By clustering the data around well-chosen objects and leveraging the triangle inequality, if a query object is sufficiently distant from any one object in the cluster, the entire cluster may be pruned. MoBIoS exploits methods of metric-space indexing where such clustering is done hierarchically. Consequently, index structures are programmatically similar to the tree-based index structures commonly used in database management systems. Similarity queries can be executed in O(log n) time.

MoBIoS is part of a general trend in building scalable sequence database index structures offline to support faster online search. Examples include the ED-tree, BLAT, SST, SSAHA, and CAFE as well as the application of suffix-trees (Tan et al., 2003; Kent, 2002; Giladi et al., 2002; Ning et al., 2001; Williams and Zobel, 2002). These efforts report online retrieval times from 6 to 200+ times faster than BLAST implementations. Of these index structures, SST, algorithmically closest to MoBIoS, partitions each sequence into overlapping k-mers and maps them to a metric vector space indexed by tree structured vector quantization (Giladi et al., 2002). CAFE uses inverted index for overlapping k-mers (Williams and Zobel, 2002). BLAT and SSAHA use hashing to index overlapping and non-overlapping k-mers respectively (Kent, 2002; Ning et al., 2001).

MoBIoS is distinguished from these other efforts by means of its generality. Each of these other efforts is a turnkey solution to a specific problem and exploits hard coding of the similarity function. Although they effectively make the case for sophisticated index structures, each software system has a tightly defined application. By integrating with the standardized architecture of database management systems, MoBIoS is not only flexible—the knowledge and effort required to customize it is typical of commodity database applications.

3 ALGORITHMS AND IMPLEMENTATION

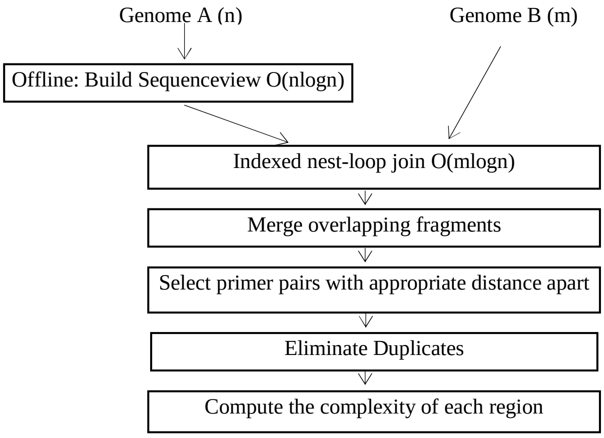
Genome A (n)
Genome B (m)
Offline: Build Sequenceview O(nlogn)
Indexed nest-loop join O(mlogn)
Merge overlapping fragments
Select primer pairs with appropriate distance apart
Eliminate Duplicates
Compute the complexity of each region

Fig. 3. Overview of finding the conserved primer pairs between two genomes

The MoBIoS storage manager is complete and has typical cursor-based interfaces. This study was done by determining a query plan likely to be generated by a SQL query engine (Figure-3) and coding and integrating the operators by hand. A full SQL front-end is under development and briefly described below.

3.1 MoBIoS index structure and search strategy

The cornerstone of this system is that, by casting biological criteria for similarity as metrics, fast algorithms may be applied. In this application we are interested in modeling the criteria for primers as a metric. The nucleotide sequence of a primer does not have to be a perfect complement to the amplicon. A small number of mismatches may occur and the PCR reaction will still function. Hence, *matching* 18-mers means the sequences are identical or may have 1 or 2 nucleotides mismatching. Given two sequences of the same length the number of pair-wise mismatches forms a metric. For a binary alphabet this is equivalent to Hamming distance.

In preliminary research we determined that of the many metric-space indexing algorithms, multi-vantage point (MVP) trees are the most promising for sequence retrieval (Miranker, et al., 2004). MVP trees enable us to store a set of sequences as 18-mers. Given an arbitrary 18-mer, it is possible to query the tree, in O(log n) time, for all matching 18-mers. In other words there exists a fast retrieval mechanism, which entails biological criteria. In related work we have succeeded in developing a metric that closely approximates PAM evolutionary distance among peptide k-mers (Xu and Miranker, 2004).

The first step is to initialize the database and instruct it to build an index on overlapping 18-mers of genome A. An MVP tree can be initialized in O(nlogn).

3.2 Genome join

The second step is to determine all *matching* pairs of 18-mers, one from each genome. Since the internal representation of the sequences is equivalent to two tables, one for each genome, where each row is an 18-mer, this comparison is equivalent to a database join (O'Neil, 1994). A simple implementation, called nested loops, involves using two loops to iterate through all possible pairings of the rows, testing each pair for a match. This algorithm executes in O(mn) time. In algorithmic terms, this is equivalent to a solution that initializes BLAST with one genome, and then executes BLAST queries on a stream of k-mers from the second genome.

A faster algorithm is called indexed-nested loops. The outer loop remains, sequentially accessing rows. The inner-loop is replaced with an index retrieval, replacing the O(n) execution of the loop with an O(logn) operation. In net, matching rows can be determined in O(mlogn). We queried the MVP tree for the k-mer and its reverse complement.

3.3 Additional steps

The whole genome sequence contains exons, introns, and intergenic regions, as well as many low-complexity regions. The low complexity regions are often referred to as subsequences with simple patterns, such as k-mer repeats, and cannot be used in primer design. Our test using 1.8mb rice sequences against 20mb Arabidopsis sequences (with about 1% low complexity regions) yielded 4,943,000 matching fragments which could generate 19,582,038 possible pairs. Intuitively, a minimum representation of a low-complexity subsequence will be smaller than that of a complex subsequence. Recall that a compression algorithm is used to find the minimum representation of a string. We implemented the LZW compression algorithm to check the complexity of a fragment by its compression ratio (Welch, 1984). With such a large volume of data, filtering out all of the low-complexity fragments using a compression algorithm takes significant additional time. Our tests showed that to remove about 1.1% of the total fragments to be searched cost a 5% time increase.

However, due to the properties of the MoBIoS index structure, fragments that are d-mer repeats can be detected by comparing the offset values among the exact matches found in search. Since the database was built by bulk loading data sequentially, the clustering process will store similar fragments in the order of their position (offset) in the original sequence. Let the set $R=\{R_1,R_2, \ldots R_n\}$ be the set of offset values that query fragment q occurs in database D. If, for any $i \in (1,n]$, $2*|R_i-R_{i-1}|=2*d \leq$ lengthOf(q), then the fragment q is composed by d-mers. Thus low complexity fragments can be eliminated without dramatically increasing overhead.

The overlapping k-mers in the results can be merged to form a set of highly conserved subsequences existing in both genomes. Subsequences with appropriate distances are selected as potential primer pairs. These primer pairs are then sorted based on their sequence complexity.

3.4 Post processing

The final step in the data analysis is to determine if putative primer pairs are conserved not only between Arabidopsis and rice, but also among other plant genomes. After filtering low complexity sequences, the remaining primer pairs and their intervening sequences were BLASTed against all of the known plant genomes in GenBank, including plant chloroplast and mitochondrial genomes. (Once a good schema mapping is determined, this step will be accomplished within MoBIoS). A matching result that contains both members of a primer pair and is not in one of the organelles is a good argument for a common evolutionary history, because the probability of finding the identical primer combination within the distance constraints is otherwise very low. It also suggests the primers will function in species other than rice and Arabidopsis. These results are being validated in a laboratory.

4 RESULTS

We detail both our biological results from comparing the Arabidopsis and rice genomes and the performance analysis of the MoBIoS core algorithms. In March 2003, the *Arabidopsis thaliana* genome and *Oryza sativa* genome were downloaded from ftp://ftp.arabidopsis.org/home/tair/ and http://www.tigr.org/tdb/e2k1/osa1/ respectively. The size of each genome is detailed in Tables 1 and 2.

Table 1. Size of Arabidopsis Genome

chromosome	1	2	3	4	5
Size (nts)	30621696	19996672	23887872	1786672	27168768

Table 2. Size of Rice Genome

Chromo-some	Size (nts)	Chromo-some	Size (nts)	Chromo-some	Size (nts)
1	65617920	5	40849408	9	31879168
2	70066176	6	47525888	10	29310976
3	43032576	7	45244416	11	30580736
4	49991680	8	46510080	12	36806656

4.1 MoBIoS' search performance

The scalability of MoBIoS is tested with Arabidopsis data. A set of databases was built ranging in size from 1 to 50 million k-mer elements. Each database was populated by a sequential load drawn from the full data set and then queried by randomly selecting 18-mers from the rice genome.

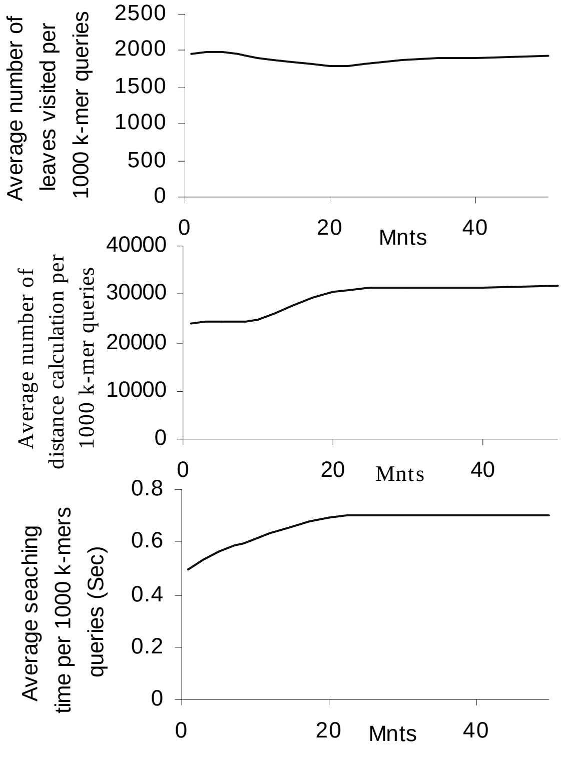

Fig. 4. The scalability test results (from top to bottom): average number of leaves visited, average number of distance calculations, and average searching time per 1000 k-mer queries vs. database.

Figure-4 illustrates, from top to bottom, the average number of distance calculations (corresponding to CPU time), the average number of leaf nodes visited (corresponding to disk I/O), and the average searching time (as wall clock time) per thousand queries for different sizes of databases. In each figure, there is no significant increase from 20M to 50M databases, indicating scalable performance. It is also interesting to note that the number of leaf nodes visited decreased slightly for a larger data set. We have reason to believe that as the database grows, the logical locality of the clusters starts to correspond better to the physical clustering (Mao et al., 2003).

4.2 Statistical results from the comparison of the Arabidopsis and rice genomes

Table-3 shows the number of conserved regions found after each step between the Arabidopsis and rice genomes. Among 13,418 possible primer pairs found from MoBIoS, about 1000 pairs had separation distances greater than 1000 bases. From these, we selected the 100 best candidates based on their complexity and queried against GenBank using BLAST. For each primer pair, the rice and Arabidopsis sequences submitted to BLAST consisted of the primers plus their separating sequences. Of these, 15 primer-pair/separating sequence combinations matched other plant genomes. Due to the abundant results, our current search radius is limited to zero, the case of exact matching. If mismatches were to be allowed, additional results are expected by increasing search radius of range queries. The problem was solved in less than 2 days with 4 concurrent processes on a multi-processor Sun 6800. We anticipate a join-operator inspired by merge-join to solve this problem in a few hours on a single processor (see section 5 for detail).

Table 3. Number of conserved regions found through comparison of Arabidopsis and rice

Number of total possible pairs	~537M X 120M
Number of low-complexity regions filtered out during search	~2.2M
Number of exact matching of 18nt fragments	~108M
Number of conserved region pairs after merge and distance filter	951,108
Number of unique conserved region pairs after pattern filter	13,418

4.3 Preliminary biological experiment results

Based upon the BLAST results, the 15 most promising primer pairs were synthesized (IDT, Inc.) and used in PCRs to assess their ability to amplify fragments in species other than rice and Arabidopsis. Six species in the sunflower genus *Helianthus* and six species in the orchid genus *Phalaenopsis* were used as templates for the PCR amplifications. *Helianthus*, like Arabidopsis, is a dicot, and *Phalaenopsis*, like rice, is a monocot, although neither of the two is particularly closely related to their taxonomic twin. These characteristics make them good test cases for the generality of the primer pairs in seed plants. PCR amplifications were performed according to standard protocols (Linder et al. 2000) using Taq polymerase. Results from amplifications were run on 1.5% agarose gels and amplified fragments were visualized using UV light and ethidium bromide staining. A primer combination was considered high quality when a single amplification band was produced in all 12 species. Of the 15 primer pairs, two produced high quality amplifications in all 12 species. This result suggests that there will be a 1-2% success rate in biological wet labs from the set of putative primers. Extrapolating to the full set of 13,418 unique conserved primer pairs, we expect between 134 and 268 biologically useful primers to be discovered that will work in the majority of seed plants.

5 AN mSQL SOLUTION TO THIS PROBLEM

Our practice with these algorithms and other related applications has been driving the development of a biological SQL extension, mSQL, which manages sequence data and includes the formal extensions to relational algebra necessary to integrate these extensions into object-relational query engines. The mSQL query in Figure 5 encodes a solution to the conserved primer pair problem. With the completion of the algebra, it is a straightforward engineering task, now underway, to build a query engine that will compile such queries to the algorithms detailed above. Thus we expect MoBIoS will soon provide a flexible environment where biologists may easily develop the solution to a wide range of genomic and proteomic problems while benefiting from, but without having to become experts in, these algorithms.

```
1.  SELECT merge(R1.fragment, A1.fragment)
2.  FROM rice_sview R1, rice_sview R2, arab_sview A1, arab_sview A2
3.  WHERE
4.  distance('base_pair_mismatch', R1.fragment, A1.fragment) ≤ 1.0
    AND
5.  distance('base_pair_mismatch', R2.fragment, A2.fragment) ≤ 1.0
    AND
6.  (FRAGOFFSET(R2.fragment)-FRAGOFFSET(R1.fragment)) ≥ 400
    AND
7.  (FRAGOFFSET(R2.fragment)-FRAGOFFSET(R1.fragment)) ≤ 3000
    AND
8.  (FRAGOFFSET(A2.fragment)-FRAGOFFSET(A1.fragment)) ≥ 400
    AND
9.  (FRAGOFFSET(A2.fragment)-FRAGOFFSET(A1.fragment)) ≤ 3000
    AND GROUP BY R1.fragment, A1.fragment
```

Fig. 5. Query for conserved primer pair discovery

We have determined that few extensions are necessary. They include a metric-join operator and a small set of operators for the decomposition of sequences into k-mers. We present these details as an explanation of the conserved primer pair query in Figure 5.

5.1 The Logical and Physical Database Structures

Before considering the query, one must first define and populate the database. Figure 6 illustrates the schema declaration for our problem. To simplify the presentation we assume that there is a single table, *genomes* (line 1), defined as follows. Each row of the database contains, as a single sequence, a representation of a chromosome, annotated by organism name and accession number (line 2). In addition to the standard SQL data types, MoBIoS includes built-in data types for DNA and Protein sequences, and Spectra (to support Mass Spectroscopy). Note that, in line 2, the chromosome sequence is defined as type DNA_sequence, and not a character or blob field.

```
1.   CREATE TABLE genomes
2.   (Organism varchar[256], Accicd varchar[20], chromsm
     DNA_Sequence);
3.   CREATE SEQUENCEVIEW rice_sview(fragment) AS
4.        SELECT CREATEFRAGMENTS(Accid, chromsm,18, 1)
5.        FROM genomes
6.        WHERE Organism = 'rice'
7.             USING base_pair_mismatch ;
8.   CREATE SEQUENCEVIEW arib_sview(fragment) AS
9.        SELECT CREATEFRAGMENTS(Accid, chomsm,18, 1)
10.       FROM genomes
11.       WHERE Organism = 'arab'
12.            USING base_pair_mismatch ;
```

Fig. 6. Physical database definition for query in Figure 5

An endemic paradox in biological sequence analysis is that storage and identification of sequences is by long functional units, but analysis and retrieval of sequences is based on matching subsequences. In this example, genomes are represented as a set of chromosomes, but the analysis stipulates that primers constitute matching substrings a minimum of 18nt long.

We address this paradox by introducing *sequenceviews*. In relational databases *views* are used to create multiple logical representations of a database (O'Neil 1994). A view definition comprises a view name and a query. The view name can be used anywhere in a SQL program that a table name can be used. Conceptually a table is *virtually* populated by the results of the query each time the view is referenced. Thus, views are often seen as a form of SQL subroutine.

Like relational views, *sequenceviews* comprise:

- Determining sequences that are to be included in this *sequenceview* in the standard method involving a SQL query.

In addition, *sequenceviews*

- Detail the decomposition of the sequences into k-mers (*createfragments).*
- Specify a [metric] index as the access-path for the k-mers.

A *sequenceview* declaration (See Figure 6, line 3) specifies a view name, rice_sview. The new primitive *createfragments* (line 4) takes as parameters an attribute name that forms a foreign key, the attribute name specifying the field containing the long sequence, the length of the k-mer, and the step size taken by a sliding window used to form the k-mers (which is almost always 1). Thus, lines 3-7 in Figure 6, specify the creation of a virtual table, called rice_sview, where the chromosome sequences of rice are decomposed into consecutive 18-mers, overlapping by 17nt. The database will build a secondary metric-space index based on *base-pair-mismatch* to quickly locate the k-mers. We create a similar *sequenceview* on Arabidopsis (line 8).

5.2 mSQL query

Given that there is now a fast access to the contents of (virtual) tables of 18-mers of the rice and Arabidopsis genomes, the conserved primer pair query in Figure 5 can be explained as follows:

1) We wish to locate sets of two pairs of 18-mers, one pair from each genome. (lines 2)
2) The sequences of the pair of rice 18-mers must match the sequences of the pair of Arabidopsis 18-mers. One base-pair mismatch is allowed. (lines 4,5)
3) The pair of rice 18-mers must be between 400 and 3000nt apart, (lines 6,7)
4) and the same must hold for Arabidopsis. (lines 8,9)

After pattern-matching it is often desirable to merge overlapping k-mers back into a longer sequence. The mSQL syntax exploits the existing keyword *group by*, to specify that the results of the query comprised of overlapping k-mers should be operated on as disjoint sets. This is similar to the use of *group by* in the relational setting to collect rows that have an identical value in an attribute. We introduce a new aggregate function *merge* that takes said groups and concatenates the overlapping k-mers. *merge* has subtleties beyond the scope of this paper and will be formally treated elsewhere (Miranker et al., 2004).

5) For the final result the overlapping minimal length sequences are merged to give long, putatively conserved primer regions. (Lines 1 and 9).

The complete calculation presented in Section 4 also considers matching reverse complements and filtering low complexity regions. These too can be accomplished with only modest embellishment of the program of Figures 5 and 6, but including their implementation would go beyond the space limits of this paper.

6 DISCUSSION AND FUTURE WORK

We have implemented language and physical structures that will enable database management systems to directly support biological sequence analysis. By integrating biological search predicates into the query engine, we derive a single programming model for both relational and biological data, which simplifies the process of bioinformatics discovery. The MVP index structure was shown to have scalable search performance for large data sets. Its inclusion in MoBIoS will enable query engines to compile investigative queries into algorithms whose complexity includes an O(log n) term for operations now commonly done in O(n). For comparative analysis of whole genomes where n ranges from 10^6-10^9, multiple orders of magnitude performance gains are possible. Our study of the rice and Arabidopsis genomes demonstrates the feasibility of the approach.

```
SELECT merge(R.fragment, A.fragment)
FROM rice_sview R, arab_sview A
WHERE distance('base_pair_mismatch', R.fragment, A.fragment) <= 1.0
GROUP BY R.fragment, A.fragment
```

Fig. 7. SQL statement for metric join

Indeed, analysis involving the convolution of two or more complete genomes (*whole genome join*) is a problem of primary and increasing significance. Each time a new organism is sequenced it must be mapped. This means that the sequence is annotated with the location and, if possible, the function of each gene as well as a number of other important features. As the corpus increases this is more commonly being accomplished by locally aligning the entire new sequence with all previously mapped sequences and deducing that similar substrings have similar function (Gusfield, 1997). This is just the tip of the iceberg. With the availability of the data, new genomic analysis protocols requiring whole genome joins are being developed at an increasing rate (Marcotte, and Date, 2001; Rouchka, et al., 2002; Stuart, et al., 2003; Lenhard, et al., 2003).

While our results from using an indexed-nested loop metric join represent a step forward, we believe that this is only the beginning. As a database operator, a metric-space join is similar to a spatial join. The goal is to determine pairs of objects that fulfill a certain distance predicate. In Figure 7, we illustrate the SQL program for identifying q-grams that differ by at most one nucleotide. Given a tree-structured access path one can anticipate merge-join-like algorithms that would tend toward O(n) execution time, assuming output size is small. In the case of a self-join, a simple recursive pattern is effective over most, if not all, tree-based methods of metric-space indexing (Wang and Shasha, 1990). The general problem is much more difficult and remains open.

The results from the Arabidopsis and rice genome comparison produced conserved primer pairs that appear to amplify single-copy regions in a broad set of seed plants. More biological validation is underway to determine the sequences of the amplified regions and to determine (1) whether some primer pairs that did not perform well can be optimized and (2) whether our initial success rate is indicative of what will be found as more primer combinations are tried. We believe that predicting conserved regions through genomic comparison can provide valuable information for wet lab experiments to help accurately reconstruct evolutionary history.

ACKNOWLEDGEMENTS

We acknowledge Wesley Sutton and Rui Mao, students in the laboratory who have contributed to the MoBIoS project. We thank the Texas Advanced Computing Center at The University of Texas at Austin for use of the advanced computing resources. This research was supported in part by grants from NSF: DBI-0241180, IIS-0325116, EIA-0121680, EF-0331453.

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.

Bozkaya T., and Ozsoyoglu M. (1997) Distance-based indexing for high-dimensional metric spaces. *In Proc. ACM SIGMOD International Conference on Management of Data* 357-368. May11-15, 1997, Tucson, AZ U.S.A

Bozkaya, T., and Ozsoyoglu, M. (1999) Indexing Large Metric Spaces for Similarity Search Queries Association for Computing Machinery, *Transactions on Database System*, 1-34.

Chavez, E., Navarro, G., Baeza-Yates, R., and Marroquin, J.L (2001) Searching in metric spaces. *ACM Computing Surveys*. 33(3): 273-321.

Giladi, E., Walker, G. M., Wang, J.Z., and Volkmuth, W. (2002) SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*. 18(6): 873-879.

Grant, V. (1971) *Plant Speciation.1st ed.* Columbia University Press, New York, U.S.A

Gusfield, D. (1997) *Algorithms on Strings, Trees and Sequences Computer Science and Computational Biology*. Press Syndicate of the University of Cambridge, USA, 449-454.

Hillier, L. and Green, P. (1991) OSP: a computer program for choosing PCR and DNA sequencing primers. *PCR Methods* Appl., 1, 124-128

Kent, W. J. (2002) BLAT-The BLAST like alignment tool, *Genome Res*. 12, 656-664.

Lenhard, B., Sandelin A., Mendoza1 L., PEngstrm1, Jareborg N., and Wasserman W. W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* 2:13.

Li, P., Kupfer, K.C., Davies, C.J., Burbee, D., Evans, G.A. and Garner, H.R (1997) PRIMO: a primer design program that applies base quality statistics for automatd large-scale DNA sequencing *Genomics* 40 476-485.

Linder, C. R., Goertzen, L. R., Heuvel, B. V. , Francisco-ortega, J. and Jansen, R. K. (2000) The complete external transcribed spacer of 18S-26S rDNA: Amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. *Molecular Phylogenetics and Evolution* 14: 285-303.

Mao, R., Xu, W., Singh, N. and Miranker, D. P. (2003) An Assessment of a Metric Space Database Index to Support Sequence Homology. *In the proceeding of the 3rd IEEE Symposium on Bioinformatics and Bioengineering,* Mar. 10-12, Washington D.C. U.S.A

Marcotte, E.M., and Date, S.V. (2001) Exploiting Big Biology: Integrating Large-scale Biological Data for Function Inference. *Briefings in Bioinformatics* 2(4): 363-374.

Miranker, D. P. Xu, W. & Mao, R. (2003) Architecture and Application of MoBIoS, a Metric-Space DBMS to Support Biological Discovery. *In the proceeding 15th International Conference on Scientific and Statistical Database Management.* 241-244, Jul. 9-11 2003. Cambridge, MA, U.S.A.

Miranker, D.P., Liu, W., Xu, W. and Mao, R. (2004) Sequence View: A Database Mechanism for Biosequences *Technical Repot # TR-04-05* University of Texas at Austin, Computer Science Department, http://www.cs.utexas.edu/users/mobios

Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res.*, 11:1725-1729

O'Neil., P., (1994) *Database: Principles Programming and Performance,* Morgan-Kaufman, San Francisco, CA, U.S.A.

Podowski, R.M. and Sonnhammer, E.L.L. (2001) MEDUSA: large scale automatic selection and visual assessment of PCR primer pairs *Bioinformatics* 17:656-657.

Rieseberg, L.H. and Linder, C.R. (1999) Hybrid classification: Insights from genetic map-based studies of experimental hybrids. *Ecology,* 80:361-370.

Rieseberg, L.H., Sinervo, B., Linder, C.R., Ungerer, M.C. and Arias, D.M. (1996) Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. *Science*, 272:714-745.

Rouchka, E. C. Gish, W., and States D. J. (2002) Comparison of whole genome assemblies of the human genome. *Nucleic Acids Res.*, 30(22): 5004 – 5014.

Schuler, G.D. *(1997)* Sequence Mapping by Electronic PCR *Genome Methods* 7:541-550.

Stuart J. M., Segal E., Koller D., and Kim S. K. (2003) A Gene Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 302: 249-55.

Tan, Z., Cao, X., Ooi, B.C., and Tung, A.K.H. (2003) The ed-tree: an index for large DNA sequence databases *In Proc. 15th International Conference on Scientific and Statistical Database Management* 151-160, Jul. 9-11 2003. Cambridge, MA, U.S.A.

Venter, J.C., et.al. (2001) The sequence of the human genome. *Science*, 291:1304-1351.

Wang, T.L., and Shasha, D. (1990) Query processing for distance metrics. *In Proc. of the 16th International Conference on Very Large Databases,* 602-613, Aug 13-16, Brisbane, Australia,

Welch, T.A. (1984) A technique for high performance data compression *IEEE Computer*, Vol. 17, No. 6, 8-19.

Williams, H.E. and Zobel, J. (2002) Indexing and Retrieval for Genomic Databases. *IEEE Transactions on Knowledge and Data Engineering,* 14(1): 63-78.

Xu, W., and Miranker, D.P. (2004) A metric model for amino acid substitution, in press *Bioinformatics,*

Xu, W., Miranker, D.P., Mao, R., and Wang, S. (2003) Indexing Protein Sequences in Metric Space. *Submitted for publication,* http://www.cs.utexas.edu/users/mobios

Yianilos, P. (1993) Data structures and algorithms for nearest neighbor search in general metric spaces. *In Proc. 4th ACM-SIAM. Symposium on Discrete Algorithms* (SODA'93) 311-321, Jan.25-27, 1993 Austin, Texas U.S.A.