gpALIGNER: A Fast Algorithm for Global Pairwise Alignment of DNA Sequences

Hadian Dehkordi, Mostafa; Masoudi-Nejad, Ali*+

Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, I.R. IRAN

Mohamad-Mouri, Morteza

Department of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, I.R. IRAN

ABSTRACT: *Bioinformatics, through the sequencing of the full genomes for many species, is increasingly relying on efficient global alignment tools exhibiting both high sensitivity and specificity. Many computational algorithms have been applied for solving the sequence alignment problem. Dynamic programming, statistical methods, approximation and heuristic algorithms are the most common methods applied to this problem. We introduce gpALIGNER, a fast pairwise DNA-DNA global alignment algorithm. gpALIGNER uses similar score schema with DIALIGN-T to produce the final alignment. It also uses the concept of "spaced seeds" to determine locally aligned subsequences which construct semi-global alignment as the preliminaries of global alignment computation. This enables gpALIGNER to have the precision provided by the DIALIGN-T algorithm in considerably less time and space complexities. We performed benchmarking of our approach based on numerous datasets from standard benchmarking databases and real sequences of NCBI database where gpALIGNER performed three times faster than DIALIGN-T. gpALIGNER is a new alternative for having sensitivity and selectivity of DIALIGN-T but with less computational cost.*

KEY WORDS: *Sequence alignment, Pairwaise alignment, Sequence comparison, Dynamic programming, Spaced seeds.*

INTRODUCTION

Global sequence alignment is one of the oldest and the most common challenging tasks in bioinformatics, which has many applications in biology. Although it has a long history in computer science, it is still an active area of research, which introduces computational challenges partially solved by existing algorithms. To understand how different species are related to each other, we need to find similarities among their DNA sequences. Many researches have been done on the analysis of differences among species through their DNA and RNA. Study of RNA and DNA sequences is achievable with global alignment algorithms. Due to the large size of DNA databases, accuracy and efficiency of the alignment algorithm became an important issue in the DNA analysis of different species; there are more demands to have an accurate and efficient algorithm for global alignment tools exhibiting both sensitivity and selectivity.

* *To whom correspondence should be addressed.*
+ *E-mail: amasoudin@ibb.ut.ac.ir*
1021-9986/11/2/139 8/$/2.80

Over the last decade, various methods have been developed for exact string matching and more importantly for approximate string matching, which have resulted in software applications that are now widely used by scientists. Some of these methods are based on the Dynamic Programming approach; *Needleman-Wunsch* [1], K-tuple methods such as BLAST [2], FASTA [3] and their improved versions; [3,4], statistical methods (Hidden Markov Models) and anchor based methods such as MUMmer [5], AVID [6], LAGAN [7] and segment based methods such as DIALIGN-2 [8] and DIALIGN-T [9].

Sequence alignment is known as an optimization problem for computer scientists, which needs a score schema and a suitable algorithm as the two major prerequisites. Underlying score schema is the most import factor in this field of research. Choosing a biologically wrong schema will result in biologically inaccurate alignments through assigning high scores to biologically wrong alignments. Obviously, it is necessary to apply biological knowledge in algorithm and methods to find more meaningful results. Using a more realistic score schema even with facile heuristics ideas may lead to biologically tenable alignments.

There are many programs, which uses long seeds approaches to construct local alignments like BLAST, FASTA, and BLAT. These approaches find short exact matches "Seed" and then extend them to longer alignments. The most recent DNA homology search software like PatternHunter [10] and YASS [11] improves BLAST power and speed with the new 'spaced seed' idea. Spaced seed technique improves sensitivity without losing sensitivity. In spaced (aka gapped or discontiguous) seeds approach instead of matching the whole string, matches of the short string at some pre-selected positions are required. Significant improvement on the sensitivity and the speed of homology searches has been achieved by optimization of these positions. This fact was also independently investigated by *Buhler et al.* [12] and *Brejova et al.* [13]. Today the spaced seed is a widely accepted practice (for reviews see B*rown et al.* 2004[14]). This innovation triggered various studies [15-19] related to the usage, design and generalizations of spaced seeds.

In this paper, we introduce a novel extension to original DIALIGN algorithm to achieve a better alignment with high sensitivity and specificity using less time and space complexities in pairwise DNA-DNA global alignment. We found that finding diagonals in sequence alignment algorithms in DIALIGN with very short seeded index approaches like chaos [7] is a time consuming job. Using fast index to search algorithms, especially with longer seeds (spaced seeds) and using biological observations to construct diagonals will improve the selectivity and sensitivity of algorithm. To find local similarities we used an algorithm to enhance the sensitivity of DNA similarity search based on *Noe*'s idea [11], which is more sensitive than BLAST and BLAT on low-scoring similarities. Moreover, one of its advantages over PatternHunter is the possibility of using transition-constrained seeds, which gives an improvement in sensitivity by 15-20% on coding and/or transition-rich regions. The algorithm also provides better and less redundant alignments compared to the REPuter [20].

METHODS AND ALGORITHMS

Most of the algorithms used for optimization problem rely on some kind of scoring scheme (objective function) to assign a quality score to every possible alignment of a given input sequence set and an optimization algorithm to find optimal or near optimal alignment based on such a scoring scheme. The score schema which was used in gpALIGNER is based on similar score schema with DIALIGN to produce the final alignment, but in preparatory steps, in order to select local alignments we made them assigned a weight to them based on blast and length algorithm of the short string (see *Subramanian et al.,* 2005[9] for further information).

FORMAL DEFINITION

A biological sequence $S = s_1...s_n$ is a finite sequence of symbols over a finite set of nucleic or amino acids alphabet $\sum$. Let $S = s_1...s_n$ and $T = t_1...t_n$ be two sequences over a finite alphabet $\sum$. A global pairwise alignment A of S and T is a pair of sequences S' and T' of symbols over alphabet $\hat{\sum} = \sum \cap \{-\}$ results from inserting gaps $\{-\}$ into both S and T such as the two augmented sequences, after the insertion of gaps, have the same length with no gap in the same column. Using the abovementioned criteria, it is obvious that there are

$$\sum_{l=\max(m,n)...m+n} \binom{1}{1-n, 1-m, m+n-1}$$

different global alignments of a sequence of length n with a sequence of length m. Obviously, the number of possible alignments will grow exponentially, but only one of them would be the best candidate for maximal global alignment. The term of "best" could be defined by a score schema, which assigned highest possible score to that candidate. Candidates with higher score are better global alignments regarding the selected score schema.

Score Schema

The score schema used in this study was originated from DIALIGN-T, which is based on similarities among whole sequences rather than similarities among single residues. First introduced by *Morgenstern et al.* [8], this score schema can be applied to both locally or globally related data sets. In the schema each fragments f is assigned a weight w(f) depending on the probability p(f) of random occurrence of such fragment. In this schema, score of an alignment -a consistent set of fragments- $A=\{f_1,...f_k\}$ is defined as sum of fragments' weight score $w(f_i)$ defined as negative logarithms of probabilities $p(f_i)$ of their random occurrence.

$$\text{Score}(A) = \sum_{f\in A} w(f) = \sum_{f\in A} -\log(P(f)) = -\log \prod_{f\in A} P(f)$$

In order to find the optimal alignment with maximal score, we need to find a collection of fragments with minimal product of $\prod_{f\in A} P(f)$

For mathematical treatment of this problem, see Morgenstern et al. [21]. However, DIALIGN-T is an improvement to the original DIALIGN and contains more heuristic improvement to reduce the influence of isolated local similarities (by excluding low-scoring sub-fragments) and adjust weight score of fragments by taking the similarity of sequences prior to the greedy procedure into account.

gpALIGNER's Algorithm

The gpALIGNER algorithm proceeds as following steps:

1- Applying a spaced seed local alignment algorithm in order to generate local alignments between the two sequences

2- Constructing a semi global alignment by chaining an ordered subset of the local alignments as fragments

3- Computing the final global alignment considering semi global alignment, actually aligning the regions between fragments by finding the best alignment that stays within a limited area around the semi global alignment

Dividing initial large alignment problem into smaller and manageable ones will save computation time. But we need to apply a sensitive local alignment algorithm in the first step and accurate procedure to construct global map in second step to avoid sub optimal solutions to the problem. In the following section, we describe each step in more details.

Generate Local Alignments

During the initial step of forming the algorithm, information about all possible seeds contained in the input sequence (s) should be collected. This can be done through a traditional procedure: given a size k, we store in a hash table the positions of all k-words occurring in the sequence. For each k-word, the hash table contains a linked list of its positions in the sequence.

After this preparatory step, the algorithm is composed of two parts. The first part is a linking algorithm. It considers seeds, i.e. repeated k-words extracted from the hash table, and processes them to form groups of seeds, according to criteria based on the distances between corresponding k-words. The second part is an extension algorithm that triggers and performs the extension of some of the constructed groups of seeds. Triggering the extension is derived from a selection criterion, called *group criterion*, based on the total nucleotide size of the group (note that seeds can overlap). Groups of seeds verifying these criterions are submitted for further extension (see Noe et al. 2005 [22]for more details). After generating a set of all local alignments ,to construct alignment we try to select a subset of these alignments, which most likely to be part of final global alignment.

Spaced Seeds

We have assumed that nucleotide mutations occur independently along similar regions. However, this assumption is not always justified. In particular, in protein-coding regions, the third codon base is more prone to mutations than the first and the second one. A way to consider this observation is to use spaced seeds. In contrast to classical seeds that correspond to k contiguous nucleotide matches, spaced seeds are represented by a shape, which specifies

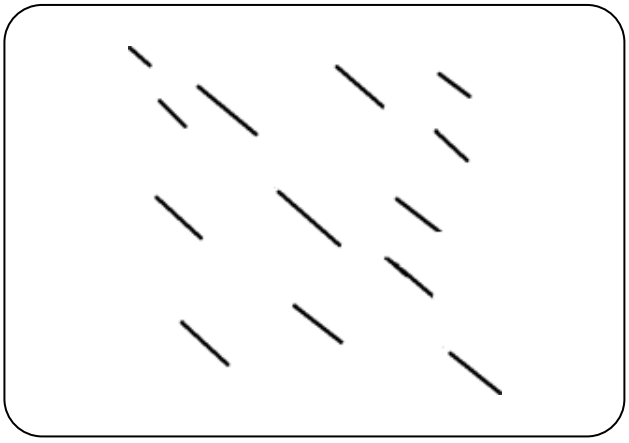

Fig. 1: Compute local alignments using spaced seeds which can be represented as diagonal in dot matrix.

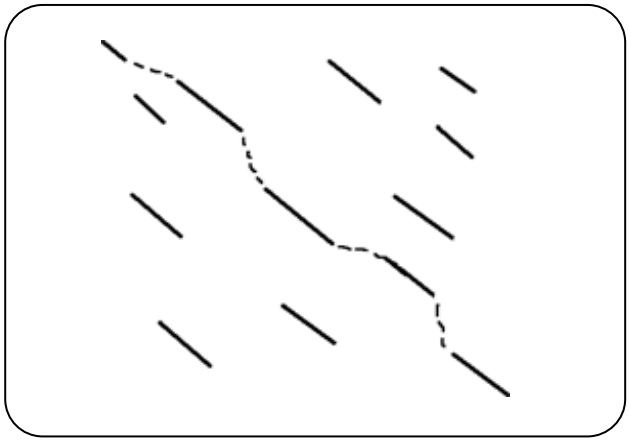

Fig. 2: Compute semi global alignment by chaining an ordered subset of the local alignments

positions at which matches occur (Fig. 1). Different models for spaced seed has been proposed; the model used in gpALIGNER is based on algorithm introduced by Noe et al. [22]. In this model seeds are sequences on $\{\#,-,@\}$ in which $\#$ represents nucleotide matches, -- represents don't care symbol and @ represents match or transition (mutation A↔C or C↔T). The weight of seed is defined as the number of # plus half of @. The choice of the shape is important and directly affects the efficiency of the seed. Spaced seeds have been shown to considerably improve the sensitivity, not only on protein coding regions but also on general unconstraint DNA sequences. Recently, spaced seeds have been designed and systematically used in many algorithms, and have been studied, from theoretical perspective [11].

Constructing Semi Global Alignment

Given a set of possible fragments, we must pick the ones we want to use as diagonals for our alignment. In many cases, this step is not trivial one since there are thousands of possible fragments of which a non-conflicting set may include only dozens. Chaining algorithm addresses this problem; the simplest problem of this sort is the longest increasing substring problem, in which we only seek the largest possible set of diagonals (Fig. 2).

Let D = $\{d_1,d_2,...,d_q\}$ be the set of maximal diagonals, largest possible sets of diagonals. In order to select diagonals we need to assign a weight to each diagonal. For each d_i in D we denote its weight $w(d_i)$ to be the product of its BLAST score and length. Moreover, for two distinct diagonals $d=(s_1,e_1,s_2,e_2)$ and $d'=(s_1',e_1',s_2',e_2')$ in the set D we define $d<d'$ if and only if the two inequalities $e_1<s_1'$ and $e_2<s_2'$ hold. In this set s_1 and s_2 represent starting positions of diagonal in sequences also e_1 and e_2 represent final positions of diagonal in sequences

A diagonal set is a collection of non-overlapping, non-crossing maximal matches, which for each d and d' the relation $d<d'$ holds. To construct semi global alignment we need to select a set of diagonals that has the largest total weight. We used a Longest Increasing Subsequences Algorithm to solve this problem [7].

Computing Global Alignment

The last step in forming our algorithm is aligning the region between diagonals selected in previous steps. Semi global alignment, diagonals discovered in previous steps, are used as the initial diagonals to DIALIGN-T algorithm. In this step, this set of diagonals might be extended and new diagonals would be discovered. We align the regions between these diagonals using original DIALIGN-T alignment algorithm and more accurate string matching algorithms such as chaos. Since the running time of program depends on the average number of diagonals in pairwise alignment, using identified diagonals in semi global alignment improves response time of the program by limiting the search space for subsequent iterations. In this case the running time of algorithm will be improved up to three times.

RESULTS AND DISCUSSION

We evaluated the performance of our program and compared it to alternative global alignment software tools

Table 1: The summary statistics of running 6 alignment algorithms on BRaliBase 2.1.

Sequences (K2)	ClustalW	gpALIGNER	Dialign-T	Dialign-2-2	Lagan	Muscle
Entero_5_CRE	0.988542	0.984583	0.983542	0.962708	0.840417	0.987917
Entero_OriR	0.961429	0.964286	0.96102	0.905714	0.946735	0.956327
HCV_SLVII	0.953333	0.963922	0.971765	0.939608	0.986078	0.968431
Retroviral_psi	0.936966	0.941348	0.938539	0.906292	0.865955	0.946629
S_Box	0.758132	0.770111	0.768571	0.74967	0.795604	0.781011
TAR	0.993357	0.983112	0.978287	0.984755	0.94542	0.993287

such as ClustalW [23] DIALIGN-T, DIALIGN2.2, LAGAN, MUSCLE [24] as a pairwise DNA-DNA global alignment using a variety of datasets. We conducted a comprehensive benchmark on the BRAliBase database version 2.1 [25]. This database is an enhanced RNA alignment benchmark for sequence alignment programs. The database was constructed using alignments from release 5.0 of the Rfam database, and provided a reasonable-sized data set of homologous RNAs of different families. Table 1 summarizes the output of BRaliBase 2.1 for ClustalW, ClustalX, DIALIGN-T, DIALIGN2-2 and LAGAN, based on k2 dataset which can be used for pairwise algorithms. The program we used for scoring (*compalignp*) is available from the BRAliBase 2.1 [25]. For comparison purposes, all programs were run on a PC machine with AthlonX2 2.2 GHz CPU with 3 GB RAM. According to the figures and table, all these programs are state of the art and their results in some cases are close to each other. If sequences are locally related gpALIGNER is superior to other global alignment algorithms. Moreover, if sequences are globally related, alignment of gpALIGNER is comparable with other sequence alignments such as ClustalW, Muscle and LAGAN.

We also compared gpALIGNER with DIALIGN-T in terms of selectivity and sensitivity. In both cases gpALIGNER outperformed DIALIGN-T. One of drawbacks in DIALIGN-T is that it usually takes relatively too much time to align large sequences, which in most cases is not practically acceptable in real world when there are a few alignment tools having slightly better quality compared to DIALIGN-T in shorter time. Our results reveal a dramatic improvement over existing alignment algorithms. Particularly, when time, space, and coverage are considered in aggregate, gpALIGNER is clearly performing better than current DIALIGN-T sequence alignment algorithm. In our approach, constructing global map helps to achieve alignment for large sequences in less time.

Most of benchmarks for sequence alignments are based on short sequences, which do not provide any information about the performance of alignment programs dealing with larger sequences, but in reality, these programs are supposed to be able to deal with very large sequences. This led us to create a collection of sequences extracted from GenBank to investigate the performance of alignment programs on longer sequences. Instead of directly measuring alignment accuracy, which is impossible when true alignments are unknown, we assessed global alignment quality by the relative length of all conserved regions aligned by algorithms. From a biological point of view, an accurate global alignment should correctly align evolutionarily related regions, including the syntenically conserved regions. Therefore, the relative length of syntenically conserved regions aligned can be used as an indirect measure for assessing the quality of global alignments by different tools.

In most of test cases, gpALIGNER is superior to DIALIGN-T. To observe the differences, we randomly picked 16 pairs of sequences from the GenBank chloroplast genomes. Table 2 contains results of running both programs on several test cases. A summary of statistical information for each test case presented. Apart from that information, output of alignment was evaluated based on a simple scoring method. In most of test cases not only conserved region aligned by gpALIGNER are larger but also the number of gaps in alignment is less and number of exact matches detected by gpALIGNER is more than those in DIALIGN-T.

Different methods have been used to define conserved regions. one of the most common method is based on

Table 2: The summary statistics of running gpALIGNER and DIALIGN-T on large sequences. Length of conserved regions identified by gpALIGNER is more than DIALIGN-T.

#	GenBank ID	Length (Kbp)	Algorithm	Score1	EM	MM	NoG	AGL	CRL	Time
1	NC_001320	134	gpALIGNER	-171528	22385	38958	58	3046	7705	14
	NC_001840	164	DIALIGN-T	-196595	17779	32645	91	2181	5473	33
2	NC_004543	161	gpALIGNER	-10610	64309	58442	244	320	42082	13
	NC_005086	162	DIALIGN-T	-45370	58517	47660	325	342	44950	39
3	NC_001840	164	gpALIGNER	-181244	31422	42702	83	2415	13562	15
	NC_006137	183	DIALIGN-T	-226509	24247	27067	176	1397	12030	46
4	NC_007407	153	gpALIGNER	27646	71131	53734	289	199	55896	9
	NC_008829	153	DIALIGN-T	11949	65663	59239	222	259	48376	32
5	NC_005973	134	gpALIGNER	233987	125629	861	40	399	125318	21
	NC_008155	134	DIALIGN-T	268462	134399	46	19	4	134539	42
6	NC_006050	159	gpALIGNER	160952	111203	37025	585	30	107400	14
	NC_009275	155	DIALIGN-T	145705	105923	42098	505	36	98520	38
7	NC_008591	136	gpALIGNER	-15610	59547	46488	254	336	44080	7
	NC_004543	161	DIALIGN-T	-38900	54824	42780	323	316	40455	34
8	NC_007407	153	gpALIGNER	159841	109798	34316	568	33	-	14
	NC_009265	154	DIALIGN-T	147445	105128	39760	484	36	-	37
9	NC_008591	136	gpALIGNER	-15610	59547	46488	254	336	44080	7
	NC_004543	161	DIALIGN-T	-38900	54824	42780	323	316	40455	34
10	NC_008591	136	gpALIGNER	-14753	57790	47172	260	308	43983	8
	NC_008829	153	DIALIGN-T	-39260	51582	46957	206	452	37340	28
11	NC_000926	121	gpALIGNER	100426	79988	40015	229	74	-	7
	NC_009573	135	DIALIGN-T	82457	73866	46184	194	87	-	23
12	NC_004561	156	gpALIGNER	-73280	45660	52805	147	749	25776	8
	NC_004766	150	DIALIGN-T	-104863	38655	49402	163	803	19856	35
13	NC_004766	150	gpALIGNER	-78865	44017	51298	142	803	24961	10
	NC_009266	154	DIALIGN-T	-117169	36385	43532	143	1012	20414	35
14	NC_009599	158	gpALIGNER	-195231	23180	33601	109	1897	10754	15
	NC_008101	161	DIALIGN-T	-216004	18895	29918	104	2141	7297	35
15	NC_004561	156	gpALIGNER	185454	117949	30407	553	25	116152	15
	NC_009269	154	DIALIGN-T	173237	113597	34748	468	30	107913	39
16	NC_008589	128	gpALIGNER	-121393	26785	36708	68	2022	-	8
	NC_009573	135	DIALIGN-T	-187040	15887	15143	115	1760	-	26

The terminology we have used in the tables and their meanings are as follows: GenBank ID: is the reference ID of sequence in NCBI database. Len(Kbp): is approximate length of sequence in kilo base pair format. Algorithm: is the name of algorithm, results in same row are related to algorithm. Score: is score of alignment based on a simple scoring schema, Score is calculated considering 2 for matches and -1 for mismatches. In this method we consider -11 for gap open penalty and -1 gap extend penalty. EM (Exact Match): is number of exact matches in alignment.MM (Mismatch): is number of mismatches in alignment. NoG (Number of Gaps): is number of gaps in alignment. AGL (Average Gap Length): is total length of gaps divided by number of gaps in alignment. CNL (Conserved Regions Length): Length of conserved regions identified by gpALIGNER and DIALIGN-T programs using VISTA (2000). Time: is total running time of algorithm in Minutes.

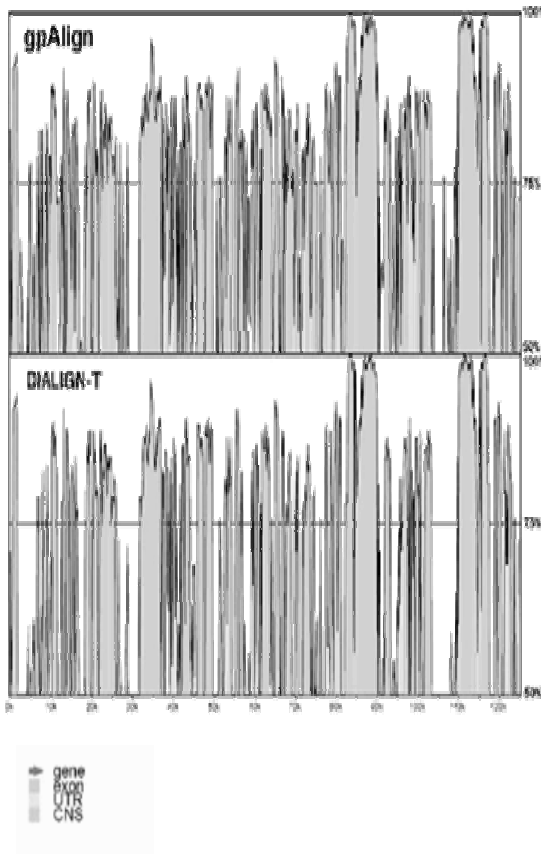

Fig. 3: Visualization of gpALIGNER and DIALIGN-T alignments on VISTA of NC_007407 and NC_008829. Conserved regions are highlighted under the curve, with red indicating a conserved non-coding region. A conserved region is defined as more than 75% identity over 100 bp stretch. gpALIGNER found significant conservation in most of regions whereas DIALIGN-T which in some regions e.g. between 5k-30 k or 90k-110 k found less conserved regions.

percentage of identifying over a region of fixed length [26,27]. We used VISTA [28], which uses this method, to extract conserved regions for each pairwise global alignment using cutoff value of 75% identity over 100 bp. An example of output generated by VISTA is given in Fig. 3. gpALIGNER not only identifies the same conserved regions but also finds more conserved regions than DIALIGN-T. To see the differences we randomly picked one sample of the table 2, comparing alignment of NC_007407 and NC_008829 as it is shown, total length of conserved regions found by gpALIGNER is more than DIALIGN-T. Nevertheless, the number of exact matches and mismatches are comparable. However, gpALIGNER achieves this supreme result 3.5 times faster.

CONCLUSIONS

In this study gpALIGNER, an improved global pairwise sequence (DNA/RNA) alignment based on DIALIGN-T alignment algorithm has been described. Its new strategy to use spaced seed to construct semi global alignment may represent an improvement over existing methods, achievingmore selectivity without loss in sensitivity. With this approach, the performance and the quality of alignment on large-scale sequences is crucially improved.

Successful experiences on using parallel algorithms into alignment algorithms show significant improvement on the running time of sequential algorithms. There are potentials to incorporate parallelism into gpALIGNER by distributing pairwise alignment or constructing semi global alignments to multiple. An upcoming release of gpALIGNER will include a parallel implementation to address multi global alignment.

Received : March 15, 2010 ; Accepted : Apr. 25, 2011

REFERENCES

[1] Needleman S.B., Wunsch C.D., A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *J. Mol. Biol*., **48**, p. 443 (1970).

[2] Gish W., Miller W., Myers E.W., Altschul S.F., Lipman D., Basic Local Alignment Search Tool, *J. Mol. Biol*., **215**, p. 403 (1990).

[3] Lipman D., Pearson W., Rapid and Sensitive Protein Similarity Searches, *Science*, **227**, p. 1435 (1985).

[4] Schaffer A., Zhang J., Zhang Z., Miller W., Altschul S., Lipman D., Gapped Blast and Psiblast: A New Generation of Protein Database Search Programs, *Nucleic Acids Res.*, **25**, p. 3389 (1997).

[5] Fleischmann S., Peterson R.D., White J., Delcher O., Kasif A.L., Salzberg S.L., Alignment of Whole Genomes, *Nucleic Acids Research*., **27**, p. 2369 (1999).

[6] Pachter L., Bray N., Dubchak I., Avid: A Global Alignment Program, *Genome Res.*, **13**, p. 97 (2003).

[7] Brudno M., Chapman M., Göttgens B., Batzoglou S., Morgenstern B., Fast and Sensitive Multiple Alignment of Large Genomic Sequences, *BMC Bioinformatics*, **4**, p. 66 (2003).

[8] Morgenstern B., DIALIGN 2: Improvement of the Segment-to-Segment Approach to Multiple Sequence Alignment, *Bioinformatics*, **15**, p. 211 (1999).

[9] Subramanian A., Weyer-Menkhoff J., Kaufmann M., Morgenstern B., DIALIGN-T: an Improved Algorithm for Segment-Based Multiple Sequence Alignment, *BMC Bioinformatics*, **6**, p. 66 (2005).

[10] MA B., TROMP J., LI M., PatternHunter: Faster and More Sensitive Homology Search, *Bioinformatics*, **18**, p. 440 (2002).

[11] Noé L., Kucherov G., Improved Hit Criteria for DNA Local Alignment, *BMC Bioinformatics*, **5**, p. 149 (2004).

[12] Buhler J., Keich U., Sun Y., Designing Seeds for Similarity Search in Genomic DNA. in "Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB'03)", 10–13 April, Berlin, Germany, ACM Press, pp. 67–75 (2003).

[13] Brejova B., Brown D., Vinar T., Vector Seeds: An Extension to Spaced Seeds Allows Substantial Improvements in Sensitivity and Specificity, "Proc. 3rd International Workshop on Algorithms in Bioinformatics (WABI'03)", LNCS, **2812**, p.39 (2003).

[14] Brown D., Li M., Ma B., Homology Search Methods. In Wong, L. (Ed.), "The Practical Bioinformatician", Singapore World Scientific Press, pp. 217–244 (2004).

[15] Brejova B., Brown D., Vinar T., Optimal Spaced Seeds for Hidden Markov Models, with Application to Homo-Logous Coding Regions, In: "R. Baeza-Yates, E. Chavez, M. Crochemore, ed., Combinatorial Pttern Matching, 14th Annual Symposium (CPM)", LNCS, **2676**, p. 42 (2003)

[16] Choi K.P., Zeng F., Zhang L., Good Spaced Seeds for Homology Search, *Bioinformatics*, **20**, p. 1053 (2004).

[17] Keich U., Li M., Ma B., Tromp J., On Spaced Seeds for Similarity Search, *Discrete Appl. Math.*, **138**, p. 253 (2004).

[18] Kucherov G., Noe´ L., Ponty Y., Estimating Seed Sensitivity on Homogeneous Alignments. In "Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE2004)", May 19–21, Taichung, Taiwan, IEEE Computer Society Press, pp. 387–394 (2004).

[19] Li M., Ma B., Kisman D., Tromp J., PatternHunter II: Highly Sensitive and Fast Homology Search, *J. Bioinform. Comput. Biol.*, **2**, p. 417 (2004).

[20] Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R., REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale. *Nucleic Acids Res.*, **29**, p. 4633 (2001).

[21] Morgenstern B., Dress A., Werner T., Multiple DNA and Protein Sequence Alignment Based on Segment-to-Segment Comparison, *Proc. Natl. Acad. Sci. USA*., **93**, p. 12098 (1996).

[22] Noé L., Kucherov G., YASS: Enhancing the Sensitivity of DNA Similarity Search, *Nucleic Acids Res.*, **33** (web-server issue):W540-W543 (2005).

[23] Thompson J.D., Higgins D.G., Gibson T.J., CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice, *Nucleic Acids Res.*, **22**, p. 4673 (1994).

[24] Edgar R.C., MUSCLE: Multiple Sequence Alignment with High Score Accuracy and High Throughput, *Nucleic Acids Res.*, **32**, p. 1792 (2004).

[25] Wilm A., Mainz I., Steger G., An Enhanced RNA Alignment Benchmark for Sequence Alignment Programs, *Algorithms. Mol. Biol.*, **1**, p. 19 (2006).

[26] Fickett J.W., Wasserman W.W., Discovery and Modeling of Transcriptional Regulatory Regions, *Curr. Opin. Biotechnol.*, **11**, p. 19 (2000).

[27] Loots G.G., Locksley R.M., Blankespoor C.M., Wang Z.E., Miller W., Rubin E.M., Frazer K.A., Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Sspecies Sequence Comparisons, *Science*, **288**, p. 136 (2000).

[28] Mayor C. et al., VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length, *Bioinformatics*, **16**, p. 1046 (2000).