Dynamic programming algorithms for discovery of antibiotic resistance in microbial genomes

Manal Helal

Centre for Infectious Diseases and Microbiology, The University of Sydney
School of Computer Science, University of New South Wales

mhelal@usyd.edu.au

Vitali Sintchenko

Centre for Infectious Diseases and Microbiology, The University of Sydney
Centre for Health Informatics, University of New South Wales

HIC'09

Introduction

❖ The emerging genome sequencing technologies and bioinformatics provide novel opportunities for studying life-threatening human pathogens and to develop new applications for the improvement of diagnosis and treatment of infections.

❖ Sequenced genomes of bacteria are doubling approximately every 20 months, however, their high-quality comparative genomic analysis depends on the availability of adequate methods [1].

❖ The existing multiple sequence alignment (MSA) methods are mostly progressive and iterative. However, most of them suffer from significant biases as they assume minimum percent identity of approximately 40% for proteins and approximately 70% for DNA sequences [2,3].

❖ The objective of this study is to test a new MSA algorithm "mmDst" and to analyse its suitability for the high-throughput comparative analysis of different microbial genomes.

❖ We have used the clinically relevant task of identifying regions that determine resistance to antibiotics to test the "mmDst" and to compare its performance with established methods.

[1] Demuth A, et al. Pathogenomics: An updated European research agenda. Infect Genet Evol 2008;8:386-93.
[2] Golubchik T, Wise MJ, Easteal S, Jermiin LS. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol 2007;24(11):2433-42.
[3] Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science 2008;319(5862): 473-6.

Definitions

❖ **Alignment of antibiotic resistance determining regions:**

    ❖ Regions of similarity or dissimilarity between a set of sequences, obtained from pathogens with known resistance or susceptibility to antibiotics (quinolones) were explored.

    ❖ Quinolone resistance has been studied extensively in many different bacterial species and is usually due to single point mutations in the target of these drugs, DNA gyrase. Resistance mutations most often occur within a stretch of 50 nucleotides, the so called "quinolone-resistance determining regions" (QRDRs), which are located in the genes for the A subunits of the enzyme gyrA gene [4].

    ❖ The resistance mutations in gyrA codons 84 or 88 usually lead to the high-level in vitro resistance but other mutations can also infrequently occur.

❖ The quality of the alignment is assessed using the following scores:

    ❖ Sum of Pairs Score is usually used to assess the performance of MSA methods. This score increases as the program succeeds in aligning more matching residues in each column in the final alignment, with minimum gap insertions all over, assuming statistical independence between columns [14].

[4] Friedberg EC, Wagner R, Radman M, Specialized DNA polymerases, cellular survival, and the genesis of mutations. Science 2002;296:1627-30.
[14] Pevsner J. Bioinformatics and functional Genomics. John Wiley, New York, 2003.

Definitions – Cont'd

❖ Assessment scores continued:

  ❖ Shannon entropy is a simple quantitative measure of uncertainty in a data set. In the context of drug resistance as conferred from single mutations, knowledge of the frequencies of different amino acids in the mutation position as drawn from resistant and sensitive populations, will enable us to guess the amino acids responsible for the resistance. This is because these amino acids were certain (low entropy) in the sensitive population, versus the uncertain (high entropy) in the resisting population [15].

❖ To identify the exact start and end of the regions of highest and lowest column scores in the alignment, a simple method was implemented. The alignment was scanned for all regions of width = 2 * the window size used in the plots of the results section. Using the sum of pairs scores generated in Table 1, every region was given a score using the following average function:

$$\text{Average Region score} = \text{sum}(c_i) / (\text{Window Size})$$

where c_i is the column sum of pairs score using the identity matrix for each column i within the region.

[15] http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_readme.html

Sources of data

❖ The set of gyrA gene sequences were extracted from following microbial genomic data available in the GenBank:

  ❖ Mycobacterium tuberculosis (NCBI Accession Number NC_000962, sequence length 2518bp);

  ❖ Mycobacterium kansasii (NCBI Accession Number Z_68207, sequence length 1648bp);

  ❖ Staphylococcus aureus MSSA476 (NCBI Accession Number NC_002953, sequence length 2665bp);

  ❖ Mycoplasma pneumoniae (NCBI Accession Number NC_000912, sequence length 2443bp);

  ❖ Clostridium difficile (NCBI Accession Number NC_009089, sequence length 2521b);

  ❖ Treponema pallidum (NCBI Accession Number NC_010741, sequence length 2428bp).

❖ Gene sequences of M.tuberculosis, M. kansasii and Staphylococcus aureus MSSA476 were grouped as quinolone susceptible.

❖ Gene sequences of Treponema pallidum, Clostridium difficile and Mycoplasma pneumoniae were classified as resistant to quinolone thus potentially harbouring changes in the gyrA gene.

High Performance new MSA Method

❖ We developed a new high performance simultaneous MSA method with the characteristics:

❖ It is based on the multi-dimensional dynamic programming optimal algorithm.

❖ It employs simultaneous alignment of all sequences using novel scoring recurrence:

❖ The score of one cell is the maximum value of the 2^k-1 neighbours' temporary scores, which is calculated as the sum of pair wise scores of all residues in the sequences corresponding the decremented index elements from the current cell index to the neighbour index, plus the multiplication of the gap score by the number of un-decremented index elements.

❖ The scoring tensor is partitioned to run in parallel on a computer cluster or a multi-core architecture [5,6].

❖ The mmDst method was tested on small HPC machines and one SGI Altix cluster of maximum 64 nodes. The processor scalability reduces the execution time as more processors were employed to achieve the minimal communication cost, and high data locality

Methodology

❖ The MSA is organised in the following steps:

  ❖ Multiple sequence alignment of the first set of sequences (sensitive to antibiotics) to derive a consensus sequence, or a profile of the known behaviour.

  ❖ Align the sensitive consensus to the highest resistant sequence "Treponema pallidum" to identify major differences.

  ❖ Align a set of resisting sequences and derive their consensus sequence,

  ❖ Align the consensus of the sensitive sequences to the consensus of the resisting sequences to identify the regions of similarity and dissimilarity (visually or calculated from the scores) of both profiles.

❖ The mmDst method was compared to existing MSA heuristic methods such as CLUSTAL W [7], MUSCLE [8], TCoffee [9], Kalign [10] and MAFFT [11], using the EBI web site [12]. These methods are based on pair-wise alignments, which are proven to be less sensitive than simultaneous alignments [13].

[7] Thompson JD. CLUSTAL W, improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Retrieved from http://wwwbimas.cit.nih.gov/clustalw/clustalw.html.
[8] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 2004;32(5):1792-97.
[9] O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. J Mol Biology 2004;340:385-95.
[10] Lassmann T, Sonnhammer ELL. Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. Nucleic Acids Res 2006;34: W596-W599.
[11] Kazutaka K, Hiroyuki T. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 2008;9:212.
[12] http://www.ebi.ac.uk/Tools/sequence.html
[13] Perrey SW, Stoye J, Moulton V, Dress A W M. On simultaneous versus iterative multiple Sequence Alignment [Report]. Bielefeld, Germany: Forschungsschwerpunkt Mathematisierung, University of Bielefeld, 1997.

Methodology – Cont'd

❖ The system was implemented on a SunFire X2200 with 2xAMD Opteron quad processors of 2.3 GHz, 512 Kb L2 cache and 2 MB L3 cache on each processor, and 8GB RAM. The sequences were aligned on a reduced search space factor "Epsilon" equals 1, which represented 0.21% of the search space for the sensitive sequences and 0.19% of the resisting sequences. The following penalties were applied: gap opening = -4, gap extension = -2, mismatch score = -1, and match score = 1.

[5] Helal M, Mullin LM, Gaeta B, El-Gindy H. Multiple sequence alignment using massively parallel mathematics of arrays. In: Proceedings of the International Conference on High Performance Computing, Networking and Communication Systems (HPCNCS- 07). Orlando, FL. USA, 2007. PP. 120-7.
[6] Helal M, El-Gindy H, Mullin LM, Gaeta B. Parallelizing Optimal Multiple Sequence Alignment by Dynamic Programming. In: Proceedings of the International Symposium on Advances in Parallel and Distributed Computing Techniques (APDCT-08) held in conjunction with 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA-08), Sydney, Australia, December 10-12 2008. PP. 120-7.

Results

Table **1**: Sum-of-Pairs Sores for the alignments produced by the different methods for the Sensitive Sequences Alignment, Resistant Sequences Alignment, Sensitive Sequences Consensus Alignment with the most resisting sequence "Treponema pallidum", and Sensitive Sequences Consensus and Resistant Sequences Consensus Alignment.

	Sensitive Seq	Resistant Seq	Sen & TP	Sen & Res Cons
mmDst	339	2231	849	1582
MUSCLE	439	3216	640	1123
TCoffee	443	2881	520	1025
CLUSTAL W	222	1966	478	1469
Kalign	-1593	-716	-285	1389
MAFFT	-3647	-4712	-1670	-2114

Table **2**: Entropy value for the aligments produced by the different methods for the Sensitive Sequences Alignment, Resistant Sequences Alignment, Sensitive Sequences Consensus Alignment with the most resisting sequence "Treponema pallidum", and Sensitive Sequences Consensus and Resistant Sequences Consensus Alignment.

	Sensitive Seq	Resistant Seq	Sen & TP	Sen & Res Cons
mmDst	23869.74	27932.28	15362.36	15430.20
MUSCLE	26855.33	25144.80	16815.42	17264.06
TCoffee	27246.99	25682.06	17355.43	17797.34
CLUSTAL W	28362.00	28240.50	17753.33	18836.40
Kalign	33336.24	34849.35	21156.91	22011.64
MAFFT	37834.49	40707.34	26597.46	37938.70
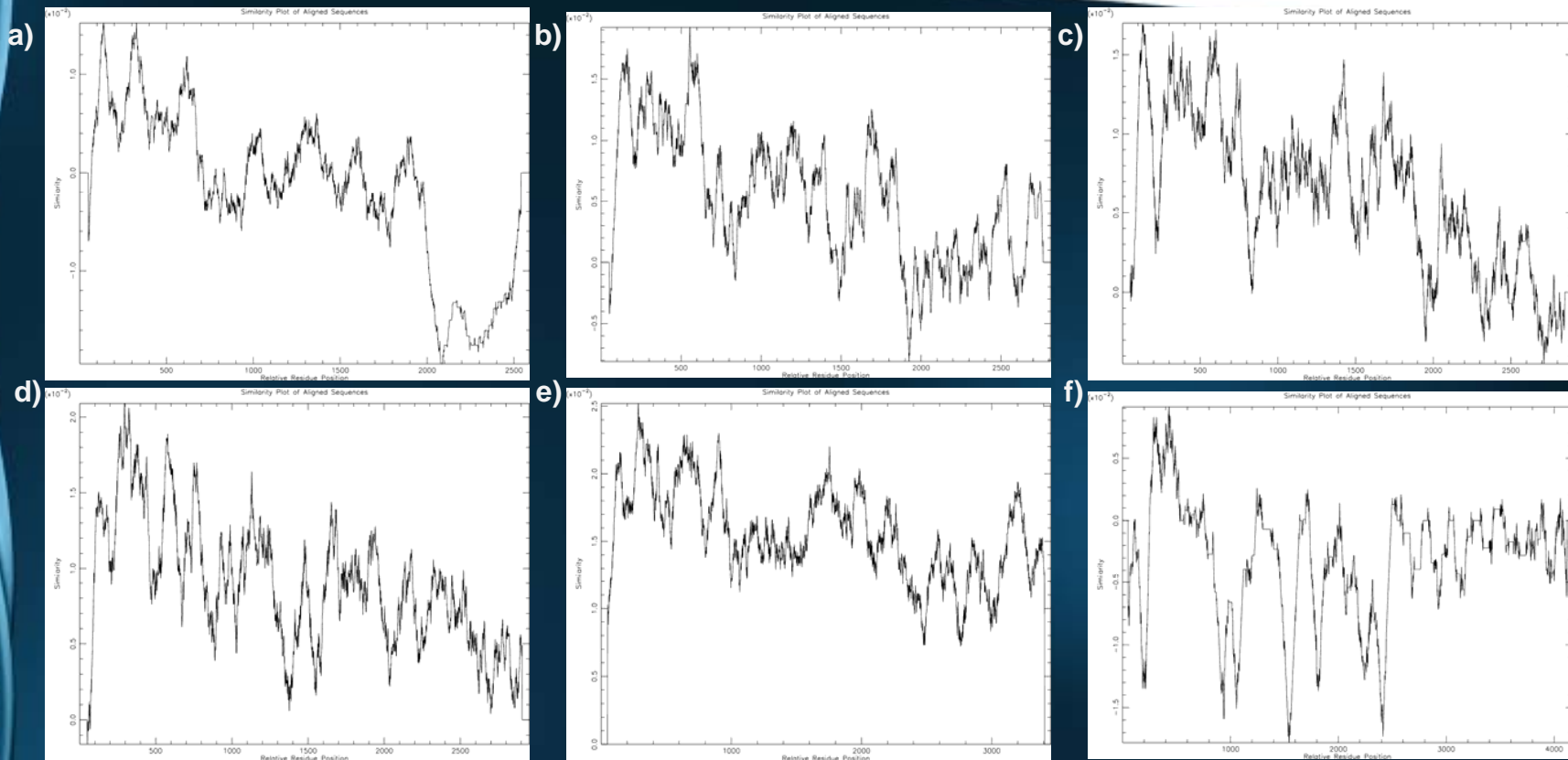
Figure 1: Similarity Regions Plot (averaged on 100 bp on the x-axis as relative residues positions) of the alignment (measured by the SP score on the y-axis) of the consensus sequence of the sensitive sequences with the most resisting sequence "Treponema pallidum" using the six different methods: a) mmDst, b) MUSCLE, c) TCoffee, d) CLUSTAL W, e) Kalign, f) MAFFT.

The similarity regions plots shown in Figure 1 are generated by plotcon tool averaged on a window size of 100 base pairs. The different alignment methods used show different areas of similarity (regions where the y-axis score is higher) and dissimilarity (regions where the y-axis score is lower), according to the SP score of the columns corresponding to the 100 base pairs averaged on the x-axis.

Results – Cont'd

Table 3: Highest and Lowest (maximum and minimum Sum-of-Pairs scores respectively) Regions (as identified in the "From" base pair number "To" base pair number) of Similarity or Dissimilarity in the alignment of the sensitive sequences consensus sequence and the "Treponema pallidum" sequence as per alignment method in the left hand side columns, and the sensitive sequences consensus sequence and resisting sequences consensus sequence alignment.

		Sensitive consensus sequence & the "Treponema pallidum" Alignment			Sensitive consensus sequence & the resisting consensus sequence alignment		
		Score	From	To	Score	From	To
mmDst	Highest	0.64	151	351	0.80	272	472
	Lowest	0.07	2167	2367	0.33	1567	1767
MUSCLE	Highest	1.03	450	650	1.11	356	556
	Lowest	-0.83	2375	2575	-0.89	2668	2868
TCoffee	Highest	1.06	233	433	1.15	430	630
	Lowest	-0.98	2495	2695	-0.36	2737	2937
CLUSTAL W	Highest	0.98	233	433	1.05	292	492
	Lowest	-0.25	2176	2376	0.18	2589	2789
Kalign	Highest	0.54	135	335	0.90	3265	3465
	Lowest	-0.54	3253	3453	-0.17	0	200
MAFFT	Highest	0.52	101	301	0.58	3915	4115
	Lowest	-1.72	2654	2854	-2.00	427	627

Conclusion

❖ The mmDst algorithm successfully identified QRDRs and handled sequences of different length. The gyrA gene of intrinsically quinolone-resistant Treponema pallidum demonstrated significant dissimilarity from gyrA genes sequences obtained from quinolone susceptible organisms of Mycobacteria (Figure 1). The details are as follows:

  ❖ Table 1 shows that "mmDst" score came third after TCoffee and MUSCLE in the first two cases, where similar sequences were aligned. Since these are progressive methods based on pair-wise alignments and building a guide tree based on an objective function, they work well with sequences of assumed similarity of 90%.

  ❖ However, in the third case where the consensus sequence of the alignment of the sensitive sequences were aligned with the most antibiotic resistant sequence which is "*Treponema pallidum*", mmDst score came second after MUSCLE.

  ❖ In the fourth case, which is the alignment of the consensus sequence of the set of sensitive sequences with the consensus sequence of the set of resisting sequences, mmDst scored the highest over all other methods.

  ❖ Aligning the profiles of both families revealed a better visual identification of the similar and dissimilar regions, rather than the alignment of one sequence representative of one family to the consensus of the other.

❖ These findings demonstrate that mmDst scores better than other heuristic methods, when aligning sequences of large dissimilarity, and can identify regions of high dissimilarity along the full length of the input sequences; a feature that can reveal more information about genes responsible for specific clinical phenotypes.