EURASIP Journal on Applied Signal Processing 2004:1, 5–12

Comparative Genomics via Wavelet Analysis for Closely Related Bacteria

Jiuzhou Song

Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1
Email: songj@ucalgary.ca

Tony Ware

Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1
Email: tware@ucalgary.ca

Shu-Lin Liu

Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1
Email: slliu@ucalgary.ca

M. Surette

Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1
Email: surette@ucalgary.ca

Received 26 February 2003; Revised 11 September 2003

Comparative genomics has been a valuable method for extracting and extrapolating genome information among closely related bacteria. The efficiency of the traditional methods is extremely influenced by the software method used. To overcome the problem here, we propose using wavelet analysis to perform comparative genomics. First, global comparison using wavelet analysis gives the difference at a quantitative level. Then local comparison using keto-excess or purine-excess plots shows precise positions of inversions, translocations, and horizontally transferred DNA fragments. We firstly found that the level of energy spectra difference is related to the similarity of bacteria strains; it could be a quantitative index to describe the similarities of genomes. The strategy is described in detail by comparisons of closely related strains: *S.typhi* CT18, *S.typhi* Ty2, *S.typhimurium* LT2, *H.pylori* 26695, and *H.pylori* J99.

Keywords and phrases: comparative genomics, gene discovery, wavelet analysis, bacterial genome.

1. INTRODUCTION

Since the publication of the whole genomic sequence of *Haemophilus influenzae* [1], the draft genomes of more than 90 bacterial strains have been completely finished. A notable outcome of these genome projects is that at least one third of the genes encoded in each genome have no known or predictable functions. The genome sequencing, while not providing the detailed minutiae of the complete sequences, allows comparisons between genomes to identify insertion, deletion, and transfers that are undoubtedly important in the different phenotype of strains. However, as the level of evolutionary conservation of microbial proteins is rather uniform, a large portion of gene products from each of the sequenced genomes has homologs in distant genomes [2]. The functions of many of these genes may be predicted by comparing the newly sequenced genomes with those of better-studied organisms. This makes comparative genomics a very powerful approach to a better understanding of the genomes and biology of the organisms and to determine what is common and what unique between different species at the genome level, especially on genome analysis and annotation. In addition, prediction of protein functions, transfer of functional information of paralogs (products of gene duplications) and orthologs (direct evolutionary counterparts), phylogenetic pattern, examination of gene (domain) fusions, analysis of conserved gene strings (operons), and reconstruction of metabolic pathways are facilitated using comparative genomics.

The large amount of data has already given rise to several studies on whole genome comparisons such as those between several closely related bacterial species [3, 4]. One problem for this kind of research is that DNA and protein fragment comparisons are highly dependent on sequence alignment methods such as FASTA34, BLAST, CLUSTALW, STADEN, PHRED, and so forth. Since the efficiency of the methods is extremely influenced by the software methods used, sequence alignment is possible for short DNA and protein sequence comparisons, the methods also need heavy use of time, energy, and resources. Here we propose a strategy for whole genome or large fragment sequence comparisons. The comparative genomics method we propose is based on the whole genome. Firstly, we use wavelet transform analysis to make a global comparison of closely related strains, giving their similarities and differences at quantitative level and with statistical meaning. Then we use keto excess or purine excess, as proposed by Freeman [5], to visualize some local differences. These indices are not like GC skew and AT skew [6, 7, 8] which depend on the sliding window size; they can show the exact positions of rearrangements and the origin and terminus sites of DNA replication. We illustrate the strategy using several closely related species including *S.Typhimurium* LT2, *S.Typhi* CT18, *S.Typhi* Ty2, *H.pylori* J99 and *H.pylori* 26695 strains. These pairs of bacteria share a similar flask-like morphology and show serological cross-reaction, but they differ in several important features including differences in G + C content and genome size, different tissue specificity, and pathogenic effects for human.

To understand the similarity between DNA structure and function, it is necessary to compare DNA sequences, especially for newly closely identified ones. Wavelet analysis has been applied to a large variety of biomedical signals; the method will provide a useful visual description of the inherent structure underlying DNA sequence [9]. A wavelet is a waveform of effectively limited duration that has an average value of zeros, and wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet [10]. It provides a multiscale representation of signals allowing efficient smoothing and/or extraction of basic components at different scales. So the wavelet analysis supplies a new way to compare whole genomes at quantitative levels. The main idea of wavelet analysis is to decompose a sequence profile into several groups of coefficients, each group containing information about features of the profile at a scale of sequence length. Coefficients at coarse scales capture gross and global features, whereas coefficients at fine scales contain the local details of the profile [11]. A wavelet variance is a decomposition of the variance of a signal; it replaces global variability with variability over scales and investigates the effects of constraints acting at different time or space scales [9]. The similarity comparison via wavelet analysis expands the traditional sequence similarity concept, which takes into account only the local pairwise DNA or amino acid sequences and disregards the information contained in coarse spatial resolution. Also the wavelet analysis does not require the complex sequence alignment processing for sequence [12]. In this study, we explore the possibility of genome comparisons using wavelet transform analysis and keto-excess or purine-excess plots to perform comparative genomics, and introduce the idea of using the energy spectra difference as a quantitative index to describe the similarity of genomes. The strategy used in this paper not only provides the location of *oriC* and *terC* sites of DNA replication, but also is a powerful tool for examining genome fragment insertion, inversion, translocation, reorganization, and revealing evolutionary history.

2. MATERIAL AND METHOD

The sequences of *Salmonella typhi* Ty2 [13], *Helicobacter pylori* J99 [14], and *Helicobacter pylori* 26695 [15] were obtained from the NCBI website; *Salmonella typhimurium* LT2 and *Salmonella typhi* CT18 were downloaded from both ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Salmonella_typhimurium_LT2/ and from ftp://ftp.sanger.ac.uk/pub/pathogens/st/, respectively.

For global comparisons of closely related bacteria, we firstly do not use sequence alignment to do the comparison, but use wavelet analysis to compare the purine-excess curve or keto-excess curve [5] and get the genome difference at quantitative level. In transforming the sequence data into digital data, we just count the cumulative number of each of the DNA bases A, C, G, and T along the whole genome. The purine excess was defined as the sum of all purines (A and G) minus the sum of all pyrimidines (T and C) encountered in a walk along the sequence up to the point plotted and was determined by

$$\text{PurineExcess}_n = \left(\sum_{i=1}^{n} B_{\mathrm{A},i} + \sum_{i=1}^{n} B_{\mathrm{G},i} - \sum_{i=1}^{n} B_{\mathrm{T},i} - \sum_{i=1}^{n} B_{\mathrm{C},i}\right), \quad (1)$$

where n ranges from 1 to N (N is the chromosome length) and $B_{\mathrm{A},i}$ is 1 if there is an A in the ith position, and 0 otherwise (the terms $B_{\mathrm{T},i}$, $B_{\mathrm{G},i}$, and $B_{\mathrm{C},i}$ are defined similarly). In the same way, the keto excess was defined as the sum of all keto bases (G and T) minus that of the amino bases (A and C) and was determined by

$$\text{KetoExcess}_n = \left(\sum_{i=1}^{n} B_{\mathrm{T},i} + \sum_{i=1}^{n} B_{\mathrm{G},i} - \sum_{i=1}^{n} B_{\mathrm{A},i} - \sum_{i=1}^{n} B_{\mathrm{C},i}\right). \quad (2)$$

Here again n ranges from 1 to N, where N is the chromosome length, and B is the number of the particular base (A, C, G, or T) occurring at the ith location (either 0 or 1 in each case). We can also define local versions of these vectors:

$$\begin{aligned} KT_n &= B_{\mathrm{T},i} + B_{\mathrm{G},i} - B_{\mathrm{A},i} - B_{\mathrm{C},i}, \\ PT_n &= B_{\mathrm{A},i} + B_{\mathrm{G},i} - B_{\mathrm{T},i} - B_{\mathrm{C},i}. \end{aligned} \quad (3)$$

The fundamental idea behind wavelet analysis is to analyze according to scale [16]. Wavelets are functions that satisfy certain mathematical requirements and are used in representing data or other functions becoming a common tool for analyzing localized variations of power within a time series, with successful applications in signal and image processing, numerical analysis, and statistics. The wavelet analysis procedure is to adopt a wavelet prototype function called an analyzing wavelet or mother wavelet. Because the original function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet function), data operations can be performed using corresponding wavelet coefficients. We employ the continuous real wavelet transform [17]. Our analyzing wavelet is the normalized first derivative of a Gaussian function:

$$\Phi(t) = \frac{t\sqrt{2}}{\pi^{1/4}\sigma\sqrt{\sigma}} \exp\left(-\frac{t^2}{\sigma^2}\right), \tag{4}$$

where σ is a scaling factor. The real wavelet transform of a function f is

$$Wf(t,s) = \int_{-\infty}^{\infty} f(u)\frac{1}{\sqrt{s}}\Phi\left(\frac{u-t}{s}\right)du. \tag{5}$$

In order to apply this transform to a vector $\underline{x}$ of length N (such as the vectors KT or PT defined above), $\underline{x}$ is taken to correspond to samples at the points $t_0 = 0$, $t_1 = 1/N$, $t = 2/N, \ldots, t_N = 1 - 1/N$ of a 1-periodic function $x(t)$. The wavelet transform Wx, for each scale s in a given range, is then just a convolution of two vectors that can be calculated in the Fourier domain using the fast Fourier transform. Explicitly, we have

$$Wx(t_i, s_j) = \sum_n x_n p_{n-i}(s_j), \tag{6}$$

where $p_i(s) = (1/\sqrt{s})\Phi(t_i/s)$, and where the sum is taken over all values n for which the terms in the sum are not negligible. The result is a two-dimensional array of values of Wx at positions t (ranging from 0 to 1) and scales s (a magnification parameter). One can think of this as a collection of one-dimensional transforms of the original signal at different scales.

Methods based on wavelet transforms generally require powerful visualization tools. In implementation, we figure out the purine excess and keto excess using Perl and C++ codes, perform wavelet transformation analysis via Matlab, and make graphics using the xmgrace graphic software on MACI-cluster parallel computers.

3. RESULTS AND ANALYSIS

3.1. Global comparison of the closely related strains

To investigate the relationship between closely related strains and determine their similarity, we use wavelet analysis to show the global spectrum of the two closely related strains. If the spectra are completely identical, they are the same strains, otherwise, we divide them to different strains. This identification, which is different from clone morphological index and physiology and biochemistry characteristics, is based on whole genome comparison. The global wavelet spectra of the purine excess for three pairs of *S.typhi* CT18 and *S.typhimurium* LT2, *S.typhi* CT18 and *S.typhi* Ty2, and *H.pylori* 26695 and *H.pylori* J99 are shown in Figures 1, 2, and 3. The power in the wavelet transform is computed for a range of scales and plotted as a function of scale level σ, where the scale is $s = 2^{-\sigma}$. The higher the scale number is, the shorter the support of the wavelet is, and so the shorter the moving window over which the signal is being measured. From Figure 1, notice the higher energy in the *S.typhi* CT18 starting at scale number 5, corresponding to a length scale of the order of 1/20 of the signal length. Using these wavelet spectra to measure the difference (in a least square sense), we find that the difference between two genomes is of the order of 1.5% of the total signal energy; the quantitative variability is also indicative of component differences in the DNA sequence. This extra variability can be observed in the cumulative signal plots for *S.typhi* CT18, in particular, in the additional features present in the signal (as compared to the corresponding graph for *S.typhimurium* LT2). From Figure 2, the lower energy in another closely-related strains *S.typhi* CT18 and *S.typhi* Ty2 energy spectra, a length scale of the order of 1/20 of the signal length, could be seen. We found that the difference between the two genomes is of the order of 0.7% difference of the total signal energy; it is definitely smaller than that between *S.typhi* CT18 and *S.typhimurium* LT2, which indicates that the similarity between *S.typhi* CT18 and *S.typhi* Ty2 is larger than that of between *S.typhi* CT18 and *S.typhimurium* LT2. From Figure 3, with a same length scale of the order of 1/20 of the signal length, the wavelet spectra measured the difference between *H.pylori* 26695 and *H.pylori* J99; the difference between the two closely related strain genomes is of the order of 17.6% of the total signal energy; it is the biggest difference in the three

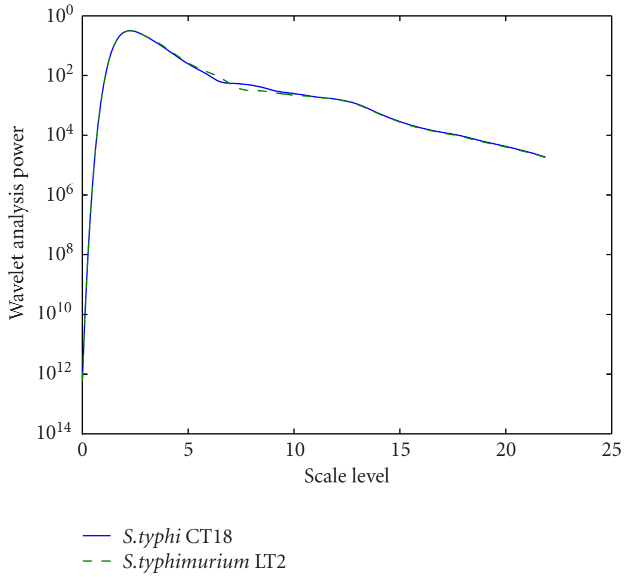


Figure 1: Comparison of the purine-excess wavelet analysis spectra in *S.typhi* CT18 and *S.typhimurium* LT2.

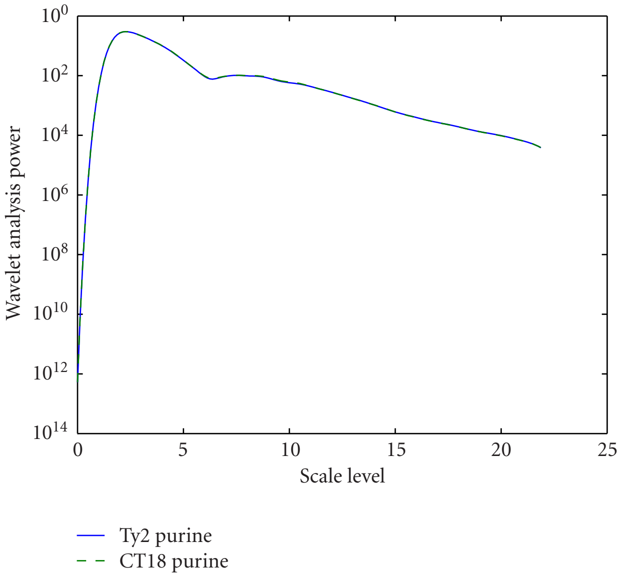


Figure 2: Comparison of the purine-excess wavelet analysis spectra in *S.typhi* CT18 and *S.typhi* Ty2.

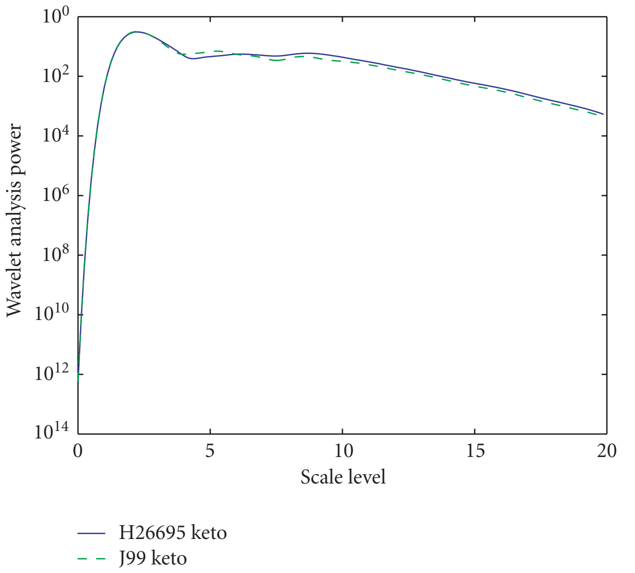


Figure 3: Comparison of the keto-excess wavelet analysis spectra in *H.pylori* 26695 and *H.pylori* J99.

compared closely related strains. Here, we can see that the variability can be observed in the cumulative signal plots for the two strains; the variability is a definite indicative of component differences in the DNA sequences. From the comparisons of the energy spectra among the strains, we can infer that the *S.typhi* CT18, compared to *S.typhimurium* LT2, has closer relationship with and bigger similarity to *S.typhi* Ty2. The strain *H.pylori* 26695 and *H.pylori* J99 have the biggest difference variability in these three compared strains.

3.2. Local comparison of the closely related strains

After comparison via wavelet transformation analysis, we have measured the global difference at a quantitative level. Now we analyze the local differences using the visualized keto-excess or purine-excess plot which explores the main information variation given by the wavelet analysis. In comparative genomics, as shown in Figure 4, the figure clearly shows the positions of terC sites and oriC sites for both strains. Most parts of the keto-excess curves overlap between *S.typhimurium* LT2 and *S.typhi* CT18, but there is an extra part around the terC site in *S.typhimurium* LT2. After partitioning in detail the fragment, the extra fragments in *S.typhimurium* LT2, the fragments A, B, C, D, E, and F in a length range from 1483934 to 1870353 bp as shown in Figure 5a, are rearranged or incompletely translocated to *S.typhi* CT18 which are also located around the terC site; the fragments are completely reversed at the length range from 1235888 to 1643129 bp and the order of fragments is reversed from fragments F to fragment A, as shown Figure 5b. The rearrangements of DNA fragments suggest that the inversions and translocations took place in the strain *S.typhi* CT18 sequences, thus disrupting the original arrangement of these fragments. As a result, the keto excess plot in the *S.typhi* CT18 is a little bit different from that of *S.typhimurium* LT2. As for the transferred or relocated genes, the most inverted fragments in *S.typhi* CT18 involve genes in *S.typhimurium* LT2 which contain cell processes: macromolecule metabolism, cell envelope, energy metabolism, such as secretion system effectors and apparatus [ssa(A–U) and yscR gene], cytoplasmic protein, inner membrane protein, family transport protein, oxidoreductase, periplasmic protein, peptide transport protein, transcriptional regulator or repression, fumarate hydratase, and tyrosine tRNA synthetase. The translocation genes in CT18 include transcriptional regulator, ATPase and phosphatase, ABC superfamily oligopeptide transport protein, peptide transport protein, anthranilate synthase, cardiolipin synthase, energy transducer, formyltetrahydrofolate hydrolase, GTP cyclohydrolase, nitrate reductase, phage shock protein, tryptophan synthase, and so forth.

Another obvious difference of the keto-excess plots in the two closely related strains is that there is a triangle peak around 4.45 mb in *S.typhi* CT18. We noted that Liu (1995) and others found that there was an insertion of length 130 kb in this region in *S.typhi* CT18. From the Keto-excess plot in Figure 4, the insertion of a large DNA fragment is confirmed. After the detailed comparison between *S.typhi* CT18 and *S.typhimurium* LT2 genomes, the insertion of a 35 kb DNA fragment ranging from 44724722 to 4507789 bp was identified in *S.typhi* CT18. DNA fragments G and H in *S.typhi* (Figure 5b) were found to be translocations from *S.typhimurium* LT2, where the fragments range from 2844714 to 2879233 bp (shown in Figure 5a). The translocation genes include regulators of late gene expression, phage

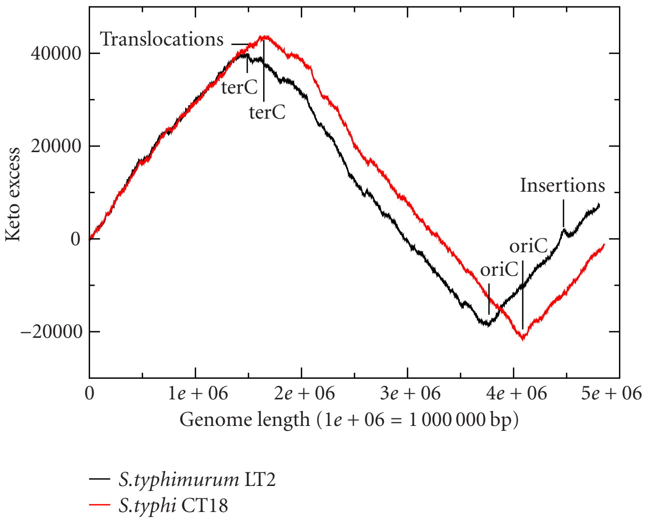
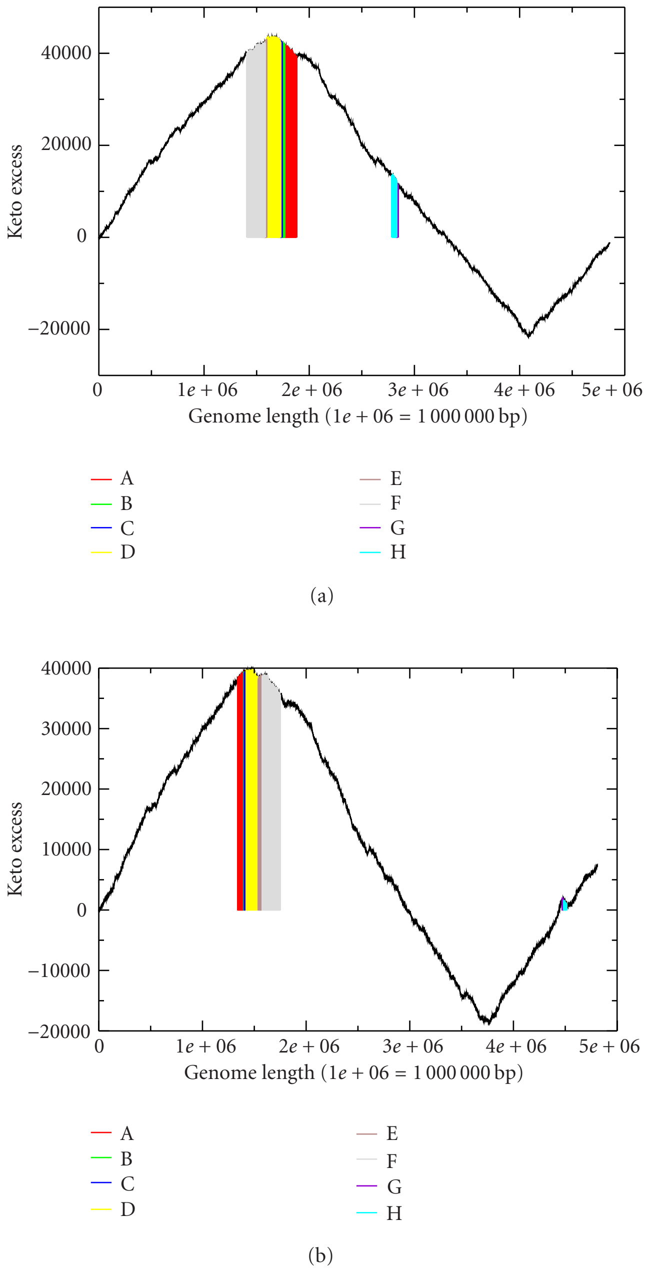


Figure 4: Comparative genomics between *S.typhi* CT18 and *S.typhimurium* LT2. The black line is keto-excess plot in *S.typhimurium* LT2 and the red one is keto-excess plot in *S.typhi* CT18. The maximum value and minimum value in each curve are corresponding to the positions of terC site and oriC site of DNA replications, respectively. Compared with *S.typhi* CT18, *S.typhimurium* LT2 has an extra part around terC site; *S.typhi* CT18 has a triangle insertion around 4.45 mb.



(a)

(b)

Figure 5: Identification of translocated and inserted fragments in *S.typhi* CT18 and *S.typhimurium* LT2. The fragments A, B, C, C, D, E, and F in *S.typhimurium* LT2 are reversed and translocated into *S.typhi* CT18; the order of fragments becomes F, E, D, C, B, A. The partial insertions in *S.typhi* CT18, fragments G and H, are horizontal transferred fragments from *S.typhimurium* LT2; the fragment length of G and H is around 35 KB.

tail protein, phage tail fiber protein, phage base plate assembly protein, lysozyme, membrane protein, and other proteins. The remaining genes within this insertion in *S.typhi* CT18 have not yet been identified.

The numbers and types of paralogs were very different between *S.typhi* CT18 and *S.typhimurium* LT2; those differences also contribute to the local differences of the wavelet transformation spectra and the keto excess-plots in the two strains. In *S.typhimurium* LT2, most of paralogs are two copies of cytochrome c-type biogenesis protein genes (ccmA-H), citrate lyase synthetase (citC-citG), and five copies of transposase (tnpA). In contrast, in *S.typhi* CT18, there are twenty-six copies of transposase (tnpA); the two copies of paralogs are oxaloacetate decarboxylase (oadA, oadB, oadG, and oadX), cytochrome c-type biogenesis protein (ccmA-H), and citrate lyase synthetase (citA-G, X, and T).

The *Salmonella enterica serovar typhi* is a human-specific pathogen causing enteric typhoid fever, a severe infection of the reiculoendothelial system. The *S.typhi* CT18 and *S.typhi* Ty2 are two well-studied pathogenic strains, by the comparison via wavelet spectra they have very little difference and are very close; this statement confirms most of researcher's inference. The information from comparative genomics and genes in *S.typhi* will help us to reveal more specific drug candidates and vaccines. Figure 6 only shows the fragments with larger than 12,000 bp. From Figure 6, the *S.typhi* Ty2 genome is distinguished from that of *S.typhi* CT18 by interreplichore inversion and translocations. The figure indicates that the inverted DNA fragments are the main reason for the difference between the two strains. There are also a lot of small inverted regions: translocated regions and unique regions (these are not shown here). Through the comparison between the strains, we found besides these major inversions that the gene structures of the two strains are very similar.

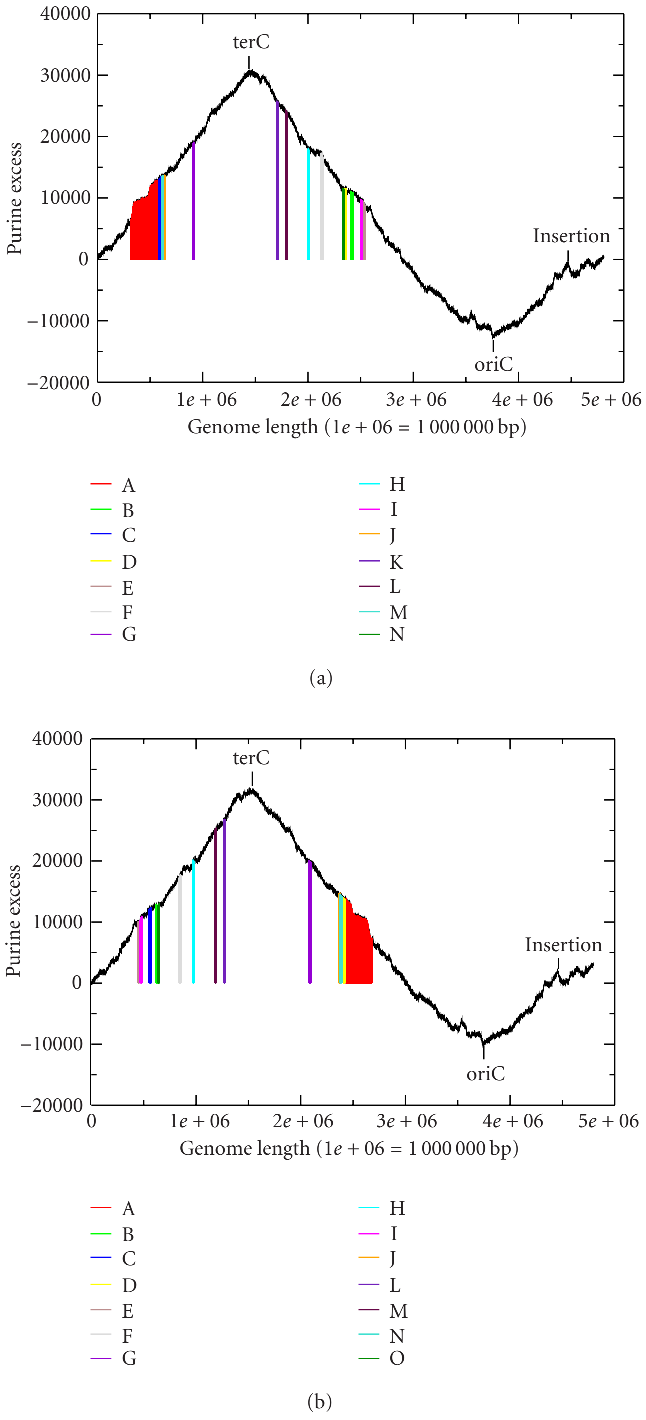


Figure 6: Identification of translocated and inserted fragments in *S.typhi* CT18(A) and *S.typhi* Ty2(B). The 14 biggest fragments A, B, C, ..., O in *S.typhi* Ty2 are reversed and translocated into *S.typhi* CT18; the order of fragments becomes O, N, M, ..., A. The partial insertions in *S.typhi* CT18 are horizontal transferred fragments into *S.typhi* Ty2; the fragment length of G and H is around 35 KB.

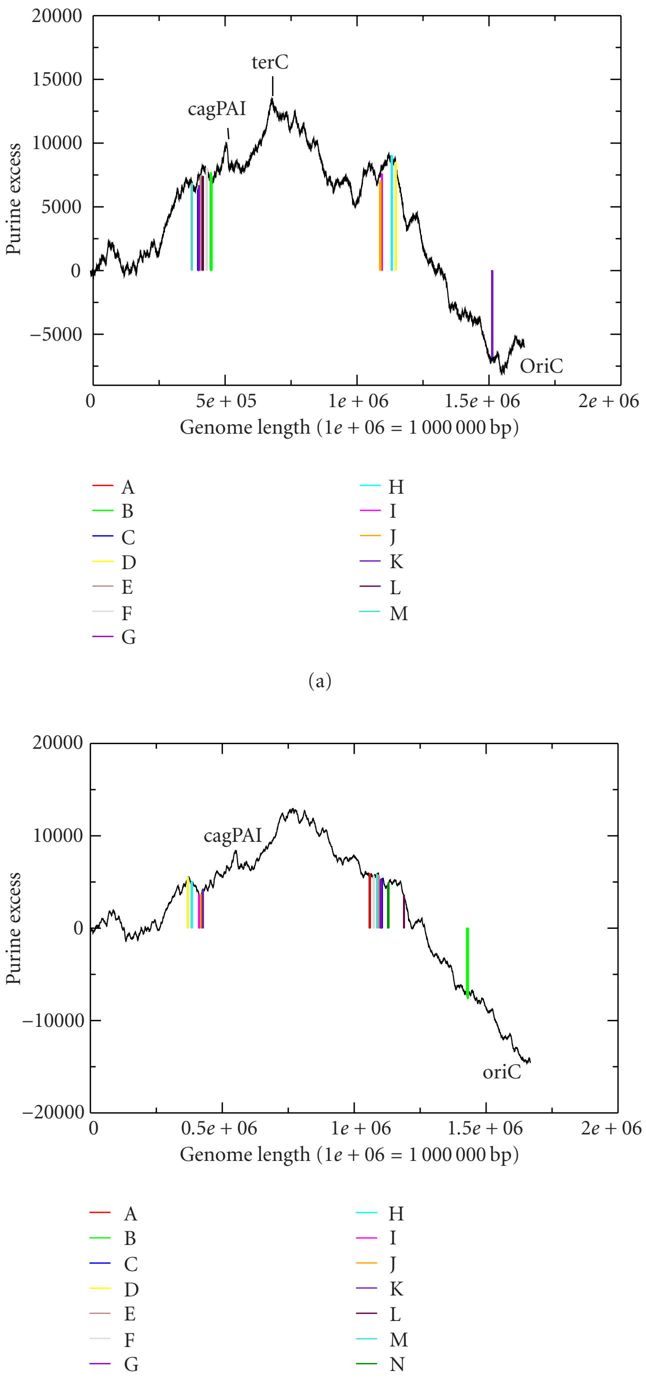
They have the same positions of oriC and terC site and physical balance features, and share a 35 kb inversions around 4.5 mb. The sequence in the inversion fragment in the two strains is the same as in the fragments G and H of the LT2. We also got a lot of pseudogenes; we think that the inverted and translocated fragments are the main reason of making the pseudogenes in the two strains. The message helps to reveal the pseudogene mechanisms and potentially contributions to pathogenicity; the detail description is beyond the scope of the paper.

Comparative genomics using purine-excess plots was also used to compare *H.pylori* strains J99 and *H.pylori* strains 26695. The size of the inversed and translocated fragments is much smaller than that of *S.typhi* CT18, *S.typhi* Ty2, and *S.typhimurium* LT2, the only fragments larger than 1000 bp are shown in Figure 7. From Figures 7a and 7b, the two strains could clearly show terC sites on the purine curves. We found that the dnaA gene is near the global minimum site, so we refer to the oriC site located on these regions. There are a lot of rearrangements, horizontal transfers, translocations, and reversions among *H.pylori* J99 and *H.pylori* 26695; the inversions and horizontally transformed DNA fragments are clearly seen to result in mirror symmetry transformations. In contrast to previous genomics comparison between the two strains, using window-sized GC skew [18], the purine-excess plots give us precise positions of inversion, translocations, and horizontal transformed DNA fragments. Interestingly, the shape and composition of cag pathogenicity island (cagPAI) are pretty similar. The inversion and translocation events do not happen in this region; this implies that the zone is not a result of differential retention of ancestral DNA in these strains but is a product of horizontal transfer; this region might represent pathogenicity islands [14]. We also found that one of the reasons which formed the jagged diagram of *H.pylori* is that *H.pylori* 26695 has some unique prologs (products of gene duplications). These prologs are acyl carrier protein (acpP), biopolymer transport protein (exbB and exbD), iron dicitrate transport protein (fecA), and transposes (tnpA and tnpB).

4. DISCUSSION

Here we have described a wavelet analysis strategy to reveal the whole genome difference between closely related bacterial strains. Compared with the widely used GC skew and AT skew, the purine excess and the keto excess are visualization tools to show whole genome information; they do not involve any default window size or the loss of any information. Via analyzing the excesses, the wavelet method enables global comparison at a quantitative level, and the keto-excess or purine-excess plot shows the local difference. Through our research, the wavelet energy spectra difference can give a quantitative measure of strain difference. It is an important value for closely related strain, especially for the similar clone morphology and serological cross reaction putative strains. It could be a quantitative index to ascertain the similarity and relationship among strains.

It is worth noting that although we can generate an enormous amount of useful information about the differences



(a)

(b)

Figure 7: Identification of translocated and inserted fragments in *H.pylori*, Strain J99 and *H.pylori*, Strain 26695. The fragments A, B, C, D, E and F in *H.pylori Strain* J99 are reversed and translocated into *H.pylori Strain* H26695.

between closely related strains or species, there is more about comparative genomic analysis other than merely identifying the presence or absence of specific fragments or genes. It is important to know whether these genes are capable of being translated into functional proteins. Very small changes such as insertion, deletion, mutation, translocations, and so forth in genomic sequence can have a disproportionate effects on the phenotype of an organism. Such changes could lead to frameshifts or base pair replacement leading to the introduction of stop codons, and may remove the activity of the encoded protein when the gene sequence is still present in the genome. In addition, these changes may produce pseudogenes. Since the changes are not random, the pseudogenes may be over-presented in certain functional classes such as pathogenicity island and cell-associated genes. For example, *S.typhi* CT18 and Ty2 contain inactivated genes which are involved in virulence and host range. For *S. typhimurium*, several genes that have been shown to be important for phenotypes in *S. typhimurium* appear to be inactive in *S.typhi* [19]. Therefore, further studies of *S.typhi* are likely to reveal rearrangements, insertions, translocations, and horizontal transfers corresponding to different tissue specificity and pathogenic effects for human and other organisms. Potentially the alteration of transcription and translation between related strains needs to be checked and confirmed by wet-bench genetic analysis. We think that although comparative genomics can provide very large amount of information on variations in each genome, it is still only an initial step in understanding the biology of an organism. Analysis of the complete genome sequence is only the start of the biological journey. The C++ and Matlab scripts for wavelet analysis and cumulative diagrams (Keto and purine excesses) are available on request from authors.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees and also Prof. C. Sensen for his comments on earlier versions of this paper. They would also like to thank Dr. Doug Phillips for his computer support.

REFERENCES

[1] R. D. Fleischmann, M. D. Adams, O. White, et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.," *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[2] E. V. Koonin and M. Y. Galperin, "Prokaryotic genomes: the emerging paradigm of genome-based microbiology," *Current Opinion in Genetics & Development*, vol. 7, no. 6, pp. 757–763, 1997.

[3] R. Himmelreich, H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann, "Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium," *Nucleic Acids Research*, vol. 25, no. 4, pp. 701–712, 1997.

[4] M. McClelland, L. Florea, K. Sanderson, et al., "Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi," *Nucleic Acids Research*, vol. 28, no. 24, pp. 4974–4986, 2000.

[5] J. M. Freeman, T. N. Plasterer, T. F. Smith, and S. C. Mohr, "Patterns of genome organization in bacteria," *Science*, vol. 279, no. 5358, pp. 1827–1829, 1998.

[6] J. R. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria," *Molecular Biology and Evolution*, vol. 13, no. 5, pp. 660–665, 1996.

[7] A. Grigoriev, "Analyzing genomes with cumulative skew diagrams," *Nucleic Acids Research*, vol. 26, no. 10, pp. 2286–2290, 1998.

[8] A. Grigoriev, "Strand-specific compositional asymmetries in double-stranded DNA viruses," *Virus Research*, vol. 60, no. 1, pp. 1–19, 1999.

[9] P. Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives," *Bioinformatics*, vol. 19, no. 1, pp. 2–9, 2003.

[10] A. Arneodo, B. Audit, E. Bacry, S. Manneville, J.-F. Muzy, and S. G. Roux, "Thermodynamics of fractal signals based on wavelet analysis: application to fully developed turbulence data and DNA sequences," *Physica A*, vol. 254, no. 1-2, pp. 24–45, 1998.

[11] J. Song, A. Ware, and S.-L. Liu, "Wavelet to predict bacterial ori and ter: a tendency towards a physical balance," *BMC Genomics*, vol. 4, no. 1, pp. 17, 2003.

[12] X.-Y. Zhang, Y.-T. Zhang, S. C. Agner, et al., "Signal processing techniques in genomic engineering," *Proceedings of the IEEE*, vol. 90, no. 12, pp. 1822–1833, 2002.

[13] W. Deng, S.-R. Liou, G. Plunkett III, et al., "Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18," *Journal of Bacteriology*, vol. 185, no. 7, pp. 2330–2337, 2003.

[14] R. A. Alm, L. S. Ling, D. T. Moir, et al., "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori," *Nature*, vol. 397, no. 6715, pp. 176–180, 1999.

[15] J.-F. Tomb, O. White, A. R. Kerlavage, et al., "The complete genome sequence of the gastric pathogen Helicobacter pylori," *Nature*, vol. 388, no. 6642, pp. 539–547, 1997.

[16] A. S. Wunenburger, A. Colin, J. Leng, A. Arneodo, and D. Roux, "Oscillating viscosity in a lyotropic lamellar phase under shear flow," *Phys. Rev. Lett.*, vol. 86, no. 7, pp. 1374–1377, 2001.

[17] S. G. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, London, UK, 1999.

[18] J. A. Abildskov, "Additions to the wavelet hypothesis of cardiac fibrillation," *Journal of Cardiovascular Electrophysiology*, vol. 5, no. 7, pp. 553–559, 1994.

[19] J. Parkhill, G. Dougan, K. D. James, et al., "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18," *Nature*, vol. 413, no. 6858, pp. 848–852, 2001.

Jiuzhou Song received his Ph.D. degree in statistical genetics from China Agricultural University in 1996. From 1996 till 1998, he held a postdoctoral fellowship in genetics at Hebrew University, and from 1998 till 2000, he was a Research Fellow in biochemistry and molecular biology at the Indiana University. Now he is a Research Associate in the Departments of Microbiology & Infectious Disease, and Biochemistry & Molecular Biology, Faculty of Medicine, University of Calgary. His main work is on bioinformatics and statistics, especially on high throughput gene expression data analysis, comparative genomics, biopathway and gene discovery, gene network, regulatory analysis, phylogenetic domain analysis, and computational biology.

Tony Ware received his Ph.D. degree in numerical analysis from Oxford University in 1991, having five years earlier obtained an honours degree in mathematics (First Class). From 1991 till 1993, he held a research fellowship in Oxford, and from 1993 till 1997, he was a Lecturer in applied mathematics at the University of Durham, UK. From 1997 till 1998, he received a research fellowship from the Department of Clinical Neurosciences at the University of Calgary. Since 2000, he has been an Assistant Professor in the Department of Mathematics and Statistics at the same university.

Shu-Lin Liu received his Ph.D. degree from Gifu University in 1990. He is an Adjunct Assistant Professor in the Department of Microbiology & Infectious Diseases, Faculty of Medicine, University of Calgary, Canada. His research focuses on bacterial evolution and speciation and is currently supported by grants from the Canadian Institutes of Health Research (CIHR) and Natural Science and Engineering Research Council of Canada.

M. Surette has been a Canada Research Chair in Microbial Gene Expression and an Alberta Heritage Foundation for Medical Research Senior Scholar since 2002. He is an Associate Professor in the Departments of Microbiology & Infectious Disease, and Biochemistry & Molecular Biology, Faculty of Medicine, University of Calgary, Canada. He has received Young Investigator Awards from Bio-Mega/Boehringer Ingelheim (Canada) in 1998–2001 and the 2000 Fisher Award from the Canadian Society of Microbiologists. His research focuses on population behaviors in bacteria and high throughput gene expression methods applied to studying bacterial virulence. His work is currently supported by grants from the Canadian Institutes of Health Research (CIHR), the Canadian Bacterial Disease Network, Genome Canada, the Human Frontiers Science Program, and Quorex Pharmaceuticals.