A multidimensional analog to the Burrows-Wheeler transform

Abhinav Nellore[1], Austin Nguyen[2], and Reid F. Thompson[3]

***Abstract*— We show how to perform multidimensional pattern matching over an n-dimensional grid of text spanning a total of s characters with nength, an analog to the Burrows-Wheeler transform. Nength exploits a Fourier duality between two kinds of grid products to map a search problem that naively takes $\mathcal{O}(s^2)$ arithmetic operations to an equivalent problem that takes $\mathcal{O}(s \log s)$ arithmetic operations.**

I. Prefix

Suppose $\mathbf{G}$ is an n-dimensional rectangular grid of numbers. Let $\mathbf{G}$ have size s_w in dimension w for $w \in \{0, 1, \ldots, n-1\}$. The total number s of entries of $\mathbf{G}$ is then given by $s = \prod_{w=0}^{n-1} s_w$. Denote the set of integers as $\mathbb{Z}$ and the set of positive integers as $\mathbb{Z}_{>0}$. Let $j \equiv_\ell k$ mean $j, k \in \mathbb{Z}$ and $\ell \in \mathbb{Z}_{>0}$ such that $j \div \ell$ gives the same positive remainder as $k \div \ell$. Write entry $(v_0, v_1, \ldots, v_{n-1})$ of $\mathbf{G}$ as $[\mathbf{G}]_{v_0 v_1 \ldots v_{n-1}}$ or equivalently $[\mathbf{G}]_{v_0, v_1, \ldots, v_{n-1}}$, and take $[\mathbf{G}]_{v_0 v_1 \ldots v_{n-1}} = [\mathbf{G}]_{t_0 t_1 \ldots t_{n-1}}$ for $v_w \equiv_{s_w} t_w$ and $w \in \{0, 1, \ldots, n-1\}$. All grids discussed in this paper are special cases of $\mathbf{G}$.

Call some grid $\mathbf{P}$ the pattern and another grid $\mathbf{T}$ the text. Define a binary operator of multiplication $\odot$ over grids with the same dimensions where, for example, $\mathbf{M} := \mathbf{P} \odot \mathbf{T}$ is given by

$$[\mathbf{M}]_{v_0 v_1 \ldots v_{n-1}} = \sum_{\{w_j\}} [\mathbf{P}]_{w_0 w_1 \ldots w_{n-1}} [\mathbf{T}]_{v_0 - w_0, v_1 - w_1, \ldots, v_{n-1} - w_{n-1}}$$

$$\text{for} \sum_{\{w_j\}} := \sum_{w_0=0}^{s_0-1} \sum_{w_1=0}^{s_1-1} \cdots \sum_{w_{n-1}=0}^{s_{n-1}-1} \text{ and } v_0, v_1 \ldots, v_{n-1} \in \mathbb{Z}. \quad (1)$$

Constrain $\mathbf{T}$ so each of its nonzero entries is an integer from the alphabet $\Omega := \{0, 1, \ldots, |\Omega| - 1\}$, where $|\Omega|$ is the size of Ω. Suppose $\mathbf{P}$ has r nonzero entries $\{p_0, p_1, \ldots, p_{r-1}\}$. Constrain $\mathbf{P}$ such that for $j, k \in \{0, 1, \ldots, r-1\}$, $p_j = |\Omega|^{r_j}$ for some $r_j \in \mathbb{Z}$, and $p_j = p_k$ if and only if $j = k$. This makes every nonzero entry of $\mathbf{P}$ a distinct power of $|\Omega|$. Consider the query grid $\mathbf{Q}$ formed by replacing every given nonzero entry p_j of $\mathbf{P}$ with some $q_j \in \Omega$ for $j \in \{0, 1, \ldots, r-1\}$, and specify a query value q as

$$q = \sum_{w=0}^{r-1} q_w |\Omega|^{r_w} \quad q_w \in \Omega. \quad (2)$$

Observe that $[\mathbf{M}]_{v_0 v_1 \ldots v_{n-1}} = q$ if and only if every nonzero entry of $\mathbf{Q}$ takes the same value at the same position in a reversed rotation $\mathbf{R}$ of $\mathbf{T}$ given by

$$[\mathbf{R}]_{t_0 t_1 \ldots t_{n-1}} = [\mathbf{T}]_{v_0 - t_0, v_1 - t_1, \ldots v_{n-1} - t_{n-1}} \quad t_0, \ldots, t_{n-1} \in \mathbb{Z}. \quad (3)$$

So imagine writing the text $\mathbf{T}$ by reversing a different grid $\mathbf{T}_r$ via

$$[\mathbf{T}]_{t_0 t_1 \ldots t_{n-1}} = [\mathbf{T}_r]_{-t_0, -t_1, \ldots, -t_{n-1}} \quad t_0, \ldots, t_{n-1} \in \mathbb{Z}. \quad (4)$$

Obtaining $\mathbf{M}$ effectively searches $\mathbf{T}_r$ for patterns matching the support of $\mathbf{P}$. Each entry of $\mathbf{M}$ uniquely determines the $\{q_w\}$ of some query $\mathbf{Q}$, so all entries of $\mathbf{M}$ with the same value are at positions in $\mathbf{M}$ that mirror the positions of all matches in $\mathbf{T}_r$ to the same query. Call $\odot$ the search product operator, and call any grid formed by multiplying grids with the search product operator a search product. In the worst case, computing $\mathbf{M}$ takes $\mathcal{O}(s^2)$ arithmetic operations: s scalar products are summed to obtain each of s entries. This paper shows how to compute $\mathbf{M}$ with $\mathcal{O}(s \log s)$ arithmetic operations.

II. Infix

The essential lesson of the Burrows-Wheeler transform (BWT) [1] is that a redundant representation of a patterned object can collapse to an economical representation that simplifies pattern matching over that object [2], [3]. Inspired by this lesson, rewrite $\mathbf{G}$ redundantly as an n-level circulant matrix $\widetilde{\mathbf{G}}$. Think of $\widetilde{\mathbf{G}}$ as a circulant matrix whose every entry is itself a circulant matrix, whose every entry is itself a circulant matrix, and so on up to a depth of $n-1$. Refer to entry $(\alpha_{n-1}, \beta_{n-1})$ of the $s_{n-1} \times s_{n-1}$ block $(\alpha_{n-2}, \beta_{n-2})$ of the $s_{n-2} \times s_{n-2}$ block $\ldots$ of the $s_1 \times s_1$ block (α_0, β_0) of the $s_0 \times s_0$ block matrix $\widetilde{\mathbf{G}}$ composed of a total of s^2 entries as $[\widetilde{\mathbf{G}}]_{(\alpha_0, \beta_0), (\alpha_1, \beta_1), \ldots, (\alpha_{n-1}, \beta_{n-1})}$. In analogy to how $\mathbf{G}$ is indexed, take $[\widetilde{\mathbf{G}}]_{(\alpha_0, \beta_0), (\alpha_1, \beta_1), \ldots, (\alpha_{n-1}, \beta_{n-1})} = [\widetilde{\mathbf{G}}]_{(\gamma_0, \eta_0), (\gamma_1, \eta_1), \ldots, (\gamma_{n-1}, \eta_{n-1})}$ for $\alpha_w \equiv_{s_w} \gamma_w$, $\beta_w \equiv_{s_w} \eta_w$, and $w \in \{0, 1, \ldots n-1\}$. Then define $\widetilde{\mathbf{G}}$ as follows:

$$\begin{aligned} &[\widetilde{\mathbf{G}}]_{(\alpha_0, \beta_0), (\alpha_1, \beta_1), \ldots, (\alpha_{n-1}, \beta_{n-1})} = [\mathbf{G}]_{t_0 t_1 \ldots t_{n-1}} \\ &\text{such that } \beta_w - \alpha_w \equiv_{s_w} t_w \\ &\text{for } w \in \{0, 1, \ldots, n-1\} \\ &\text{and } \alpha_0, \alpha_1, \ldots \alpha_{n-1}, \beta_0, \beta_1, \ldots \beta_{n-1}, t_0, t_1, \ldots, t_{n-1} \in \mathbb{Z}. \end{aligned} \quad (5)$$

[1]A. Nellore is with the Computational Biology Program, the Department of Biomedical Engineering, and the Department of Surgery at Oregon Health & Science University in Portland. anellore@gmail.com

[2]A. Nguyen is with the Computational Biology Program and the Department of Biomedical Engineering at Oregon Health & Science University in Portland. nguyenau@ohsu.edu

[3]R. Thompson is with the Computational Biology Program, the Department of Radiation Medicine, the Department of Biomedical Engineering, the Department of Medical Informatics and Clinical Epidemiology, and the Cancer Early Detection Advanced Research Center at Oregon Health & Science University in Portland as well as the Division of Hospital and Specialty Medicine in the VA Portland Healthcare System and the Portland VA Research Foundation. thompsre@ohsu.edu

Now consider a search product $\mathbf{S} := \mathbf{G}_0 \odot \mathbf{G}_1 \odot \dots \odot \mathbf{G}_{n-1}$, where the $\{\mathbf{G}_0, \mathbf{G}_1, \dots, \mathbf{G}_{m-1}\}$ are m grids. Construct the corresponding n-level circulant representations $\{\widetilde{\mathbf{G}}_0, \widetilde{\mathbf{G}}_1, \dots, \widetilde{\mathbf{G}}_{m-1}\}$ of these grids using the prescription of (5). The matrix product $\widetilde{\mathbf{S}} := \widetilde{\mathbf{G}}_0 \widetilde{\mathbf{G}}_1 \dots \widetilde{\mathbf{G}}_{m-1}$ is computed as

$$
\begin{aligned}
&[\widetilde{\mathbf{S}}]_{(\alpha_0,\beta_0),(\alpha_1,\beta_1),\dots,(\alpha_{n-1},\beta_{n-1})} \\
&= \sum_{\{\gamma_{jk}\}} \Big([\widetilde{\mathbf{G}}_0]_{(\alpha_0,\gamma_{00}),(\alpha_1,\gamma_{10}),\dots,(\alpha_{n-1},\gamma_{n-1,0})} \\
&\qquad \times [\widetilde{\mathbf{G}}_1]_{(\gamma_{00},\gamma_{01}),(\gamma_{10},\gamma_{11}),\dots,(\gamma_{n-1,0},\gamma_{n-1,1})} \\
&\qquad \times \dots \times [\widetilde{\mathbf{G}}_{m-1}]_{(\gamma_{0,m-2},\beta_0),(\gamma_{1,m-2},\beta_1),\dots,(\gamma_{n-1,m-2},\beta_{n-1})} \Big) \\
&= \sum_{\{\gamma_{jk}\}} \Big([\mathbf{G}_0]_{\gamma_{00}-\alpha_0,\gamma_{10}-\alpha_1,\dots,\gamma_{n-1,0}-\alpha_{n-1}} \\
&\qquad \times [\mathbf{G}_1]_{\gamma_{01}-\gamma_{00},\gamma_{11}-\gamma_{10},\dots,\gamma_{n-1,1}-\gamma_{n-1,0}} \\
&\qquad \times \dots \times [\mathbf{G}_{m-1}]_{\beta_0-\gamma_{0,m-2},\beta_1-\gamma_{1,m-2},\dots,\beta_{n-1}-\gamma_{n-1,m-2}} \Big) \\
&= \sum_{\{w_{jk}\}} \Big([\mathbf{G}_0]_{w_{00}w_{10}\dots w_{n-1,0}} [\mathbf{G}_1]_{w_{01}-w_{00},w_{11}-w_{10},\dots,w_{n-1,1}-w_{n-1,0}} \\
&\times \dots \times [\mathbf{G}_{m-1}]_{\beta_0-\alpha_0-w_{0,m-2},\beta_1-\alpha_1-w_{1,m-2},\dots,\beta_{n-1}-\alpha_{n-1}-w_{n-1,m-2}} \Big) \\
&= [\mathbf{M}]_{\beta_0-\alpha_0,\beta_1-\alpha_1,\dots,\beta_{n-1}-\alpha_{n-1}} \\
&\text{for } \sum_{\{\gamma_{jk}\}} := \sum_{\ell=0}^{m-2} \left(\sum_{\gamma_{0\ell}=0}^{s_0-1} \sum_{\gamma_{1\ell}=0}^{s_1-1} \dots \sum_{\gamma_{n-1,\ell}=0}^{s_{n-1}-1} \right), \\
&\quad \sum_{\{w_{jk}\}} := \sum_{\ell=0}^{m-2} \left(\sum_{w_{0\ell}=0}^{s_0-1} \sum_{w_{1\ell}=0}^{s_1-1} \dots \sum_{w_{n-1,\ell}=0}^{s_{n-1}-1} \right), \\
&\text{and } \alpha_0, \alpha_1, \dots \alpha_{n-1}, \beta_0, \beta_1, \dots \beta_{n-1} \in \mathbb{Z}. \qquad (6)
\end{aligned}
$$

Above, a change of variables is performed to substitute the dummy indices $\{w_{jk}\}$ for the dummy indices $\{\gamma_{jk}\}$. From (1) and (6), $\widetilde{\mathbf{S}}$ recovers the content of the search product $\mathbf{S}$, and in particular $\widetilde{\mathbf{M}} := \widetilde{\mathbf{P}}\widetilde{\mathbf{T}}$ recovers the content of the search product $\mathbf{M} = \mathbf{P} \odot \mathbf{T}$.

A given n-level circulant matrix $\widetilde{\mathbf{G}}$ representing the grid $\mathbf{G}$ is diagonalized by an n-dimensional discrete Fourier transform matrix $\mathbf{F}$ defined via

$$[\mathbf{F}]_{(\alpha_0,\beta_0),(\alpha_1,\beta_1),\dots,(\alpha_{n-1},\beta_{n-1})} = \prod_{w=0}^{n-1} \frac{\rho_{s_w}^{\alpha_w \beta_w}}{\sqrt{s_w}}. \qquad (7)$$

Above, the qth primitive root of unity ρ_q is given by $\rho_q = e^{-2\pi i/q}$ for $i := \sqrt{-1}$ and $q \in \mathbb{Z}_{>0}$, and $\alpha_0, \alpha_1, \dots \alpha_{n-1}, \beta_0, \beta_1, \dots \beta_{n-1} \in \mathbb{Z}$. The diagonal representation of $\widetilde{\mathbf{G}}$ is composed of at most s nonzero entries, which may be arranged in a grid $\mathbf{G}^N$ whose dimensions are the same as those of $\mathbf{G}$. Let $\{\mathbf{G}_0^N, \mathbf{G}_1^N, \dots \mathbf{G}_{m-1}^N\}$ correspond to the grids $\{\mathbf{G}_0, \mathbf{G}_1, \dots, \mathbf{G}_{m-1}\}$ in the same way $\mathbf{G}^N$ corresponds to $\mathbf{G}$. Because multiplication of diagonal matrices is equivalent to a Hadamard (entrywise) product, any search product of the $\{\mathbf{G}_0, \mathbf{G}_1, \dots, \mathbf{G}_{m-1}\}$ is dual to a Hadamard product of the $\{\mathbf{G}_0^N, \mathbf{G}_1^N, \dots \mathbf{G}_{m-1}^N\}$. Computing such a Hadamard product and transforming back to an n-level circulant representation recovers the corresponding search product.

Call the duality between search products of the $\{\mathbf{G}_0, \mathbf{G}_1, \dots, \mathbf{G}_{m-1}\}$ and Hadamard products of the $\{\mathbf{G}_0^N, \mathbf{G}_1^N, \dots \mathbf{G}_{m-1}^N\}$ nength duality. Call $\mathbf{G}^N$ a nength of $\mathbf{G}$, and say that $\mathbf{G}^N$ is obtained by nengthening $\mathbf{G}$.

Consider the Hadamard product $\mathbf{M}^N := \mathbf{P}^N \circ \mathbf{T}^N$, where $\mathbf{P}^N$ and $\mathbf{T}^N$ are the nengths of $\mathbf{P}$ and $\mathbf{T}$, respectively. This product requires at most s arithmetic operations, considerably fewer than the s^2 operations required by naive computation of the search product $\mathbf{P} \odot \mathbf{T}$. Obtaining a nength takes $\mathcal{O}(s \log s)$ arithmetic operations via a fast Fourier transform [4], and so does transforming $\mathbf{M}^N$ back to a circulant representation. So nength maps a naively $\mathcal{O}(s^2)$ search operation to an $\mathcal{O}(s \log s)$ operation.

III. Suffix

Reflect on the string case, where $n = 1$, to see how nength is an analog to the BWT. Recall that a string is encoded as integers in $\mathbf{T}$, and write all rotations of the string as the single-level circulant matrix $\widetilde{\mathbf{T}}$. But instead of pursuing a lexical sort of the rows of this matrix and peeling off the BWT from the last column of the result, diagonalize $\widetilde{\mathbf{T}}$ with a discrete Fourier transform matrix and peel off the diagonal of the result to form a nength. A lexical sort makes sense only in one dimension, and its replacement with a discrete Fourier transform is apropos: the BWT's sort maps periodic behavior in a string to runs of the same character in a transformed string, while a Fourier transform maps periodic behavior in a signal to peaks in Fourier space.

It has been 25 years since the BWT was published [1], and this paper is our paean to it. We are excited to see where extensions of the BWT will be in another 25 years.

Acknowledgements

We thank Aleksey Cherman, Ben Langmead, Craig T. Parsons, and Vanessa Vanzieleghem for their helpful feedback on early drafts of this manuscript. Abhi Nellore thanks Samia M. Hamed and Paul T. Spellman for helpful discussions. He also thanks Rachel A. Ward and the Oden Institute for Computational Engineering and Sciences at The University of Texas at Austin for hosting him as a visiting scholar August-October 2016, when he began work on extensions of the BWT.

References

[1] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Systems Research Center, Tech. Rep., 1994.

[2] P. Ferragina and G. Manzini, "Opportunistic data structures with applications," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, IEEE, 2000, pp. 390–398.

[3] ——, "Indexing compressed text," *Journal of the ACM (JACM)*, vol. 52, no. 4, pp. 552–581, 2005.

[4] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.