Young Tableaux for Gene Expressions

Sanju Vaidya (Joshi), *Mercy College*

***Abstract*: Young tableaux are certain tabular arrangements of integers. Alfred Young introduced them to describe irreducible representations of the symmetric group at the end of the 19th century. We will use combinatorial algorithms of permutations and Young tableaux to describe a modification of the research method of Ahnert et al for identifying significant genes in the biological processes studied in microarray experiments. In the last decade, DNA microarrays (DNA chips) have been used to study gene expressions in many diseases such as cancer and diabetes. To analyze data of microarray expression curves of genes, Ahnert et al associated permutations to the data points of the microarray curves. Using Monte-Carlo simulation they established bounds corresponding to various maps of permutations for any microarray curve's algorithmic compressibility which measures its significance in the underlying biological process. Using the Robinson- Schensted-Knuth procedure, we will associate Young tableaux to permutations corresponding to the data points of microarray curves. We will calculate the bound of Ahnert et al corresponding to the map which gives the length of the longest increasing or decreasing subsequence of a permutation.**

Index Terms: Algorithmic Compressibility, Hook Length, Microarray Curve, Young Tableaux

I. INTRODUCTION

In the last twenty years, Bioinformatics has brought together mathematicians, computer scientists, and biologists to analyze biological data such as nucleic acid (DNA/RNA) and protein sequences. Studying various functions of genes is important in analysis of genetic diseases. Many scientists have used DNA microarray (DNA chip) technology to study gene expressions in various diseases including and most notably, cancer ([6], [14], [13], [12]), but also diabetes ([17], [18], and [11]). DNA microarrays consist of glass slides or membranes onto which sequences of many genes are attached at fixed locations. They offer an efficient method of gathering data about expression levels (amount of mRNA produced in the cell) of various genes under different conditions. In the analysis of such data, the goal is to identify the genes which are important in the underlying biological process.

Willbrand et al [21] found a new method of identifying significant genes in microarray expression curves. For each gene, they constructed a plot of expression level as a function of progression such as a function of time or severity of disease. Depending upon the consecutive data points as increasing (positive) or decreasing (negative), they associated an up-down signature, a string of pluses and minuses, to the expression curve of each gene. Their method is based on analysis of probabilities of up-down signatures. In 2006, Ahnert et al [5] generalized the method of Willbrand et al [21] by using concepts in the field of algorithmic information theory. They computed various bounds on any microarray curve's algorithmic compressibility, which measures its significance in the underlying biological process. In order to do this, they introduced a two- step procedure for any microarray curve for a gene. In the first step of the procedure, they associated a rank permutation to the data points of the given microarray curve. In the second step, they chose a simple map γ, which acts upon the rank permutation and gives as its output a real number. In their analyses, using Monte Carlo Simulation, they established bounds corresponding to various maps of permutations for any microarray curve's algorithmic compressibility. For example, they found bounds corresponding to the maps γ_{long} and γ_{+-} which gives respectively the length of the longest increasing or decreasing subsequence of a permutation and the number of permutations with the same pattern of rises and falls. In this paper, we will focus on the bound corresponding to the map γ_{long} for the algorithmic compressibility. Moreover, in Vaidya (Joshi) [20], we computed the bound of Ahnert et al

This work was supported in part by National Institute of Health, Extramural Associates Program in 2008-2009

DOI: 10.5176/2251-3043_3.1.227

[5] corresponding to the map γ_{+-};we also computed the probabilities of up-down signatures of microarray curves defined by Willbrand et al [21]. For this we used Foulkes'[8] method for enumeration of permutations with prescribed up-down sequences and the hook-length formula of Frame et al [7]. The research methods of Willbrand et al [21] and Ahnert et al [5] have several advantages and are powerful tools to analyze large microarray data.

In this paper, we will describe a modification of the research method of Ahnert et al [5] for identifying genes which are significant in the underlying biological process. We will use some combinatorial algorithms of Young tableaux and permutations to calculate the bound of Ahnert et al [5] corresponding to the map γ_{long} for the algorithmic compressibility of a given microarray curve. Young tableaux are certain tabular arrangements of integers. Alfred Young [22, 23] used these tableaux in his studies of irreducible representations of the symmetric group S_n which is the group of all permutations of the set of integers from 1 to n. Knuth [10] refined a sorting procedure for sequences of integers, which was originated by Robinson [15] and Schensted [16]. Robinson - Schensted - Knuth procedure gives a one-to-one correspondence between permutations of the set $\{1,2,3,\cdots n\}$ and pairs of Young tableaux of the same shape. It may be noted that in Abhyankar-Joshi [1], [2], [3], [4] (Joshi is the maiden name of the author), they established many correspondences between multitableaux and multimonomials by generalizing the Robinson-Schnstead-Knuth (RSK) procedure in various ways. Additionally, Vaidya (Joshi) [19] gives a brief preview of Abhyankar-Joshi [1], [2], [3], and [4]. The RSK (Robinson [15], Schensted [16], Knuth [9]) correspondence is based on the procedures of inserting a positive integer in a standard Young tableau and deleting a positive integer from it. In this paper, we will use the RSK correspondence to associate pairs of Young tableaux to permutations corresponding to data series of genes. To calculate the bound of Ahnert et al [5] corresponding to the map γ_{long} for the algorithmic compressibility of a given microarray curve of a gene, we will use the Schensted's algorithm [16] for computing the number of permutations having certain lengths of the longest increasing and decreasing subsequences and the hook-length formula of Frame et al [7] for computing the number of Young tableaux of a given shape.

In Section III, we will describe the research method of Ahnert et al [5] for finding significant genes in the underlying biological process. In Section IV, we will review some theorems about Young tableaux and permutations. These theorems are proved in Part I of Schensted [16] and Frame et al [7]. In Section V, we will describe our modified research method of finding significant genes. Finally, in Section VI we will have discussion and conclusion. We will show that we can calculate the bound of Ahnert et al [5] on the algorithmic compressibility using the RSK correspondence between permutations and the Young tableaux and the hook length formula of Frame et al [7] instead of using Monte Carlo simulation. It is fascinating to see how combinational algorithms are useful in analyzing data of gene expressions and finding significant genes for target diseases. The analysis of gene expressions opens the door to improved diagnoses, individualized medical treatment, and earlier detections of diseases.

II. NOTATION & TERMINOLOGY

We will use the notation and terminology introduced in Ahnert et al [5] and Schensted [16].

III. RESEARCH METHOD OF AHNERT ET AL

In this Section, we will describe the research method of Ahnert et al [5] for determining the significant genes for the underlying biological process.

Research Method:

(1) Convert all microarray curves into their rank permutations. For example, a curve f of five data points with values 0.23, 0.54, 0.33, 0.78, 0.91 would be translated into the sequence 1, 3, 2, 4, 5, as 0.23 is the lowest data point, 0.54 is the third lowest point, 0.33 the second lowest, etc.

(2) Choose a simple map γ which acts upon a permutation and gives as its output a real number.

Permutations which are associated to the same number are grouped together.

(3) By using Monte Carlo Simulation, compute the value

$$K_\gamma(f) = -\log_2 p(f) - \log_2 N_\gamma$$

which is a bound on the algorithmic compressibility of microarray curve f. If the number $K_\gamma(f)$ is positive, then the gene corresponding to the curve f is significant to the underlying biological process.

In Ahnert et al [5], they used many simple maps from set of permutations to the set of all real numbers. For example, they used the map γ_{long} which gives the length of the longest increasing or decreasing subsequence of a permutation. As said in Ahnert et al [5], the number $K_\gamma(f)$ measures the significance of a given microarray curve f in a relation to the underlying variable of the series. For example, if the microarray curve f is a time series of measurement of gene expression across the duration of a cell cycle and $K_\gamma(f) > 0$, then the microarray curve f is more likely related to the cell cycle than others. This research method of Ahnert et al [5] has many advantages and is a powerful tool to analyze large microarray data.

IV. REVIEW OF YOUNG TABLEAUX

In this Section, we will review some results about Young tableaux and permutations. These results are proved in Schensted [16] and Frame et al [7].

The following lemma from Schensted [16] gives a one-to-one correspondence between the permutations and pairs of standard tableaux. The correspondence is called Robinson–Schensted –Knuth correspondence.

Lemma (4.1). There is a one-to-one correspondence between sequences made with the n distinct integers $x_1, x_2, \cdots, x_n$ and ordered pairs of standard tableaux of the same shape-the first containing $x_1, x_2, \cdots, x_n$ and the second containing $1, 2, \cdots, n$.

In the above lemma the standard tableau containing $x_1, x_2, \cdots, x_n$ is called the P-symbol and tableau containing $1, 2, \cdots, n$ is called the Q-symbol. The following Theorems (4.2 and 4.3) give relationships between the longest increasing and decreasing subsequences of a permutation and the number of columns and rows of the P-symbol. They are proved in Schensted [16].

Theorem (4.2): The number of columns in the P-symbol (or the Q symbol) is equal to the length of the longest increasing subsequence of the corresponding sequence.

Theorem (4.3): The number of rows in the P-symbol (or the Q symbol) is equal to the length of the longest decreasing subsequence of the corresponding sequence.

The following Theorem of Schensted [16] gives relationship between longest increasing and decreasing subsequences of a permutation and standard tableaux with certain shapes.

Theorem (4.4): The number of sequences consisting of the distinct numbers, $x_1, x_2, \cdots, x_n$ and having a longest increasing subsequence of length α and a longest decreasing subsequence of length β, is the sum of the squares of the numbers of standard tableaux with shapes having α columns and β rows.

The following Theorem of Frame et al [7] gives the number of standard tableaux of a given shape.

Theorem (4.5). The number of standard tableaux of a given shape containing the integers $1, 2, \cdots n$ is

$$\frac{n!}{\prod_{j=1}^{n} h_j}$$

where for $1 \le j \le n$, the number h_j is the hook length of the element j.

V. MODIFICATION OF THE RESEARCH METHOD OF AHNERT ET AL

In this section, we will describe our modification of the research method of Ahnert et al [5]. We will need the following Theorem to calculate the bound on the algorithmic compressibility, which measures the significance of the genes.

Theorem (5.1): Let f be a given microarray curve of n data points, where n is a positive integer. Let $\sigma(f)$ be the rank permutation associated to the n data points.. Let $\gamma_{long}(\sigma(f))$ = the length of the longest increasing or decreasing subsequence of the permutation $\sigma(f)$. Let S_n be the set of all permutations of the set $\{1,2,\cdots,n\}$. Let $H = \{\mu \in S_n : \gamma_{long}(\mu) = \gamma_{long}(\sigma(f))\}$. For each positive integer m, let N^m be the set of all m ordered tuples of positive integers. Let

$$A = \{\lambda = (\lambda_1, \lambda_2, \cdots \lambda_m) \in N^m :$$
$$\lambda_1 + \lambda_2 + \cdots \lambda_m = n, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m\}$$
$$B = \{\lambda = (\lambda_1, \lambda_2, \cdots \lambda_m) \in A:$$
$$\lambda_1 = \gamma_{long}(\sigma(f)), 1 \leq m \leq \gamma_{long}(\sigma(f))\}$$
$$C = \{\lambda = (\lambda_1, \lambda_2, \cdots \lambda_m) \in A:$$
$$\lambda_1 < \gamma_{long}(\sigma(f)), m = \gamma_{long}(\sigma(f))\}$$
$$K = B \cup C$$

For each $\lambda \in K$, let $n_\lambda =$ number of standard tableaux of shape λ containing the integers $1, 2, 3 \cdots n$. Then we have the following.

(1) The cardinality of the set H = $\sum_{\lambda \in K} (n_\lambda)^2$

(2) For each $\lambda \in K$, $n_\lambda = \dfrac{n!}{\prod_{j=1}^{n} h_j}$

where for $1 \leq j \leq n$, the number h_j is the hook length of the element j.

(3) $p(f)$ = $(\sum_{\lambda \in K} (n_\lambda)^2)$ / n!

Proof: For any permutation $\mu \in S_n$, we note that $\mu \in H$ if and only if the length of the longest increasing subsequence of μ = $\gamma_{long}(\sigma(f))$ and the length of the longest decreasing subsequence of $\mu \leq \gamma_{long}(\sigma(f))$ or the length of the longest decreasing subsequence of μ = $\gamma_{long}(\sigma(f))$ and the length of the longest increasing subsequence of $\mu < \gamma_{long}(\sigma(f))$. Then (1) follows from Lemma (4.1) and Theorems (4.2), (4.3), and (4.4) Clearly (2) follows from Theorem (4.5) and (3) follows from (1).

Now we will describe our modified research method step by step and explain it for five data points.

Modified Research Method:

(1) Assign a permutation $\sigma(f)$ to the given microarray curve f of n data points as described in Ahnert et al [5].

(2) Find the value $\gamma_{long}(\sigma(f))$ = the length of the longest increasing or decreasing subsequence of the permutation $\sigma(f)$.

(3) Use Theorem (5.1) to find the probability $p(f)$ of a random curve having the same γ_{long} value that f has and compute the value

$$K_{\gamma_{long}}(f) = -\log_2 p(f) - \log_2 N_{\gamma_{long}}$$

If this number is positive then gene corresponding to the curve *f* is significant to underlying biological process.

Example: Suppose the given microarray curve f has the following data points:-

0.41, 0.52, 0.31, 0.20, 0.11.

(1) The permutation $\sigma(f)$ is 4, 5, 3, 2, 1.

(2) Length of the longest increasing or decreasing subsequences of the permutation $\sigma(f)$ = 4.

(3) We will use Theorem (5.1) to calculate the probability $p(f)$. Let $H = \{\mu \in S_5 : \gamma_{long}(\mu) = 4\}$

To find the cardinality of the set H, we consider the shapes of standard Young tableaux that have 4 rows or 4 columns.

By Theorem (4.5), the number of standard tableaux of these shapes will be

$\frac{5!}{(1.2.3.5.1)} = 4$ and $\frac{5!}{(1.5.3.2.1)} = 4$. So by theorem 5.1, the cardinality of the set H = $4^2 + 4^2 = 32$. Consequently, $p(f) = 32/120.$ Clearly, $N_{\gamma_{long}} = 3$ and we could compute $K_{\gamma_{long}}(f) = -\log_2 p(f) - \log_2 N_{\gamma_{long}}$ This number is positive. So the gene corresponding to curve f is significant to the underlying biological process.

VI. Discussion and Conclusion

Many genetic diseases are caused by deletions, duplications, and rearrangements of chromosomal regions. Studying which genes are active in different types of tissues helps scientists to analyze various genetic diseases. In the last decade, many scientists used DNA microarray (DNA Chip) technology to identify genes which play a crucial role in the underlying biological process.

Ahnert et al [5] found a new method for finding significant genes in microarray experiments. Using Monte Carlo Simulation, they computed various bounds on any microarray curve's algorithmic compressibility, which measures its significance in the underlying biological process. Their method involves associating permutations to the microarray curves and computing various simple maps of permutations to real numbers. This method has several advantages. For example, it is unbiased towards any pattern in the data series.

In this paper, we described a modification of the research method of Ahnert et al [5] for identifying genes which are significant to the underlying biological process. We computed the bound corresponding to the map γ_{long} (which gives the length of the longest increasing or decreasing subsequence of a permutation) of Ahnert et al [5] on algorithmic compressibility using algorithms of Young tableaux and permutations. There is no loss of information since the Robison-Schensted-Knuth procedure gives a one-to-one correspondence between Standard Young tableaux and permutations corresponding to the data points of the microarray curves of genes. The method is also not biased toward any anticipated pattern. The hook length formula of Frame et al [7] gives a precise way of computing number of tableaux of a certain shape. Consequently, using Schensted's algorithm [16], we could calculate the bound corresponding to the map γ_{long} of Ahnert et al [5] on the algorithmic compressibility for a given microarray curve of a gene, instead of using Monte Carlo Simulation. Thus, we could determine significance of genes using combinatorial algorithms of Young tableaux. This is simply amazing!

REFERENCES

[1] Abhyankar S. S. and Joshi S. B., "Generalized coinsertion and standard multitableaux", journal of Statistical planning and Inference 34 (1993), 5-18, North-Holland.

[2] Abhyankar S. S. and Joshi S. B., "Generalized rodeletive correspondence between multitableaux and multimonomials", Discrete Mathematics 93 (1991), 1 – 17 North-Holland.

[3] Abhyankar S. S. and Joshi S. B., "Generalized roinsertive correspondence between multitableaux and multimonomials", Discrete Mathematics 90 (1991), 111 – 135, North-Holland.

[4] Abhyankar S. S. and Joshi S. B., "Generalized codeletion and standard multitableaux", Montreal Conference Proceedings, Group Actions and Invariant Theory, Canadian Mathematical Society 10 (1989).

[5] Ahnert S.E. , Wilbrand K. , Brown F.C.S. , and Fink T.M.A. (2006), "Unbiased pattern detection in microarray data series" Bioinformatics, Volume 22, 1471-1476.

[6] Chakrabarti R., Robles L.D., Gibson J. , Muroski M. (2002), "Profiling of deferential expression of messenger RNA in normal, benign, and metastatic prostate cell lines" Cancer Genet Cytogenet, 139,115-25.

[7] Frame J. S. , Robinson G. de B. , and Thrall R. M. , "The hook graphs of the symmetric group", Can. J. Math., 6 (1954), 316-324.

[8] Foulkes H. O., "Enumeration of permutations with prescribed up-down and inversion sequences", Discrete Mathematics, 15(1976), 235-252, North-Holland

[9] Knuth D. E., "Permutations, matrices, and generalized Young tableaux", Pacific Journal of Mathematics, 34 (1970), 709 – 727.

[10] Knuth D.E., "The Art of Computer Programming", volume 3, Sorting and Searching, Addison – Wesley, Reading, Massachusetts, 1973.

[11] Loring J.F., Wen X., Lee J .M., Sellhamer J., Somogyi R. (2001), "A gene expression Profile of Alzheimer's disease", DNA Cell Biol, 20, 683 – 695.

[12] Luo J.H. et al (2002), "Gene expression analysis of prostate cancers" Mol. Carcinog, 33, 25-35.

[13] Lyons- Weiler J., Patel S. , Bhattacharya S. (2003); "A classification based machine learning approach for the analysis of genome-wide expression data", Genome Res, 13, 503-12.

[14] Ramaswamy S., Ross K.N., Lander E. S., Golub T. R. (2003), "A molecular signature of metastasis in primary solid tumors", Nat. Genet, 33, 49-54.

[15] Robinson G. DEB, " On the representations of the symmetric group", American Journal of Mathematics, 60 (1938), 746 – 760.

[16] Schensted C. (1961), "Longest Increasing and Decreasing Subsequences", Canadian Journal of Math, 13, 179-191.

[17] Susztak et al (2003), "Genomic strategies for diabetic nephropathy, J Am Soc Nephrol, 14(suppl 3) S271-8.

[18] Urbanowska T., Mangialaio S., Hartmann C., Legay F. (2003), "Development of protein microarray technology to monitor biomarkers of rheumatoid arthritis disease" Cell Biol Toxicol, 19, 189-202.

[19] Vaidya (Joshi) S., "Correspondences between tableaux and monomials", Proceedings of the conference, "Algebraic Geometry and its Applications", edited by C. Bajaj, Computer Science, Purdue University, Springer-Verlag, New York (1994), 261 – 281.

[20] Vaidya (Joshi) S, "Up-down sequences of permutations for gene expressions" – forthcoming

[21] Willbrand K., Radvanyi F., Nadal J. P., Thiery J. P., Fink T. M. (2005), "Identifying genes from up-down properties of microarray expression series", Bioinformatics, Volume 21, 3859- 3864.

[22] Young Alfred, "On Quantitative Substitutional Analysis", Volume II, Proc. London Math Soc Ser, 1 35 (1902), 361 – 397.

[23] Young Alfred, "On Quantitative Substitutional Analysis", Volume III, Proc. London Math Soc Ser 2 28(1928), 255 – 292.

Sanju Vaidya (Joshi) received Ph.D in Mathematics from Purdue University, West Lafayette, IN, USA in 1989. She is an Associate Professor of Mathematics in Mercy College, Dobbs Ferry, NY, USA. Her current research interests are enumerative combinatorics and computational biology.