Up-Down Sequences of Permutations for Gene Expressions

Sanju Vaidya (Joshi), *Mercy College*

***Abstract*: We will describe modifications of the research methods of Willbrand et al and Ahnert et al for identifying significant genes in the biological processes studied in microarray experiments. Willbrand et al introduced a new method of identifying significant genes by analyzing probabilities of up-down signatures of microarray expression curves of genes. Ahnert et al generalized the method of Willbrand et al and established various bounds on any microarray curve's algorithmic compressibility which measures its significance in underlying biological process. We will compute the probabilities of up-down signatures of microarray curves defined by Willbrand et al by using Foulkes' method for enumeration of permutations with prescribed up-down sequences and the hook length formula of Frame et al. Moreover, we will compute the bound of Ahnert et al corresponding to the map which gives the number of permutations with the same pattern of rises and falls for any microarray curve's algorithmic compressibility. It is fascinating to see that how combinatorial algorithms of permutations and Young tableaux are useful in analyzing data of gene expressions and identifying significant genes in biological processes.**

Index Terms: Algorithmic Compressibility, Hook length, Microarray curve, Young tableaux

I. INTRODUCTION

Various mutations in genes can lead to serious diseases. In the last decade, many scientists have used DNA microarrays (DNA chips) to study gene expressions in various diseases such as cancer, diabetes, arthritis, and Alzheimer's disease ([6], [14], [13], [12], [17], [18],[11]). A DNA microarray is a tool that consists of a small membrane or a glass slide containing samples of many genes. In the past scientists were able to do genetic analysis for only a few genes at once. DNA microarrays allow scientists to monitor gene expression values for thousands of genes in one experiment quickly and efficiently. Microarray experiments generate massive amount of data. Various computational methods are required to analyze such data and identify genes which are significant to 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 the up-down signatures of the expression curves of genes. 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_{+-} and γ_{long} , which respectively gives the number of permutations with the same pattern of rises and falls and the length of the longest increasing or decreasing subsequence of a permutation. The research methods of Wilbrand et al [21] and Ahnert et al [5] are powerful tools to analyze large microarray data.

In this paper, we will describe modifications of the research methods of Wilbrand et al [21] and Ahnert et al [5] for identifying genes which play an important role in the underlying biological process. Our modified research methods will use

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

DOI: 10.5176/2251-3043_3.1.228

some combinatorial algorithms of Young tableaux and permutations. 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. The hook-length formula of Frame et al [7] computes the number of standard Young tableaux of a given shape and gives a combinatorial way of finding the dimensions of the irreducible representations of the symmetric group. Foulkes [8] described a method for enumeration of permutations of $\{1,2,3\cdots n\}$ with a prescribed up-down sequence. This enumeration method of Foulkes [8[is based on the dimensions of the irreducible representations of the symmetric group given by the hook-length formula of Frame et al [7] and the coefficients in the product of Schur functions. We will use Foulkes' method [8] to calculate the probabilities, given in Wilbrand et al [21], of the up-down signatures of microarray curves of genes. Additionally, we will use Foulkes' method [8] to compute the bound of Ahnert et al [5] corresponding to the map γ_{+-} for the algorithmic compressibility of a given microarray curve of a gene. It may be noted that Vaidya(Joshi)[20], computed the bound of Ahnert et al [5] corresponding to the map γ_{long} which gives the length of the longest increasing or decreasing subsequence of a permutation using the Robinson-Schensted-Knuth (RSK)[9,10,15,16] correspondence between permutations and Young Tableaux and the hook length formula of Frame et al [7]. Further it may also be noted that Abhyankar-Joshi (Joshi is the maiden name of the author) [1,2,3,4] generalized RSK correspondence and established various correspondences between multi tableaux and multimonomials and Vaidya (Joshi) [19] gave a summary of them.

In Section III, we will describe the research methods of Wilbrand et al[21] and Ahnert et al [5] for finding significant genes in the underlying biological process. In Section IV, we will review Foulkes' [8] method of enumeration of permutations with a prescribed up-down sequence and the hook length formula of Frame et al [7].In Section V, we will describe our modifications of research methods of Wilbrand et al [21] and Ahnert et al [5] for finding significant genes. Finally, in Section VI we will have discussion and conclusion. 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 is crucial in earlier detections of diseases and their treatments.

II. NOTATION & TERMINOLOGY

We will use the notation and terminology introduced in Wilbrand et al [21], Ahnert et al [5], Foulkes' [8] and Schensted[16].

III. RESEARCH METHODS OF WILBRAND ET AL & ANHERT ET AL

In this Section, we will review the research methods of Wilbrand et al [21] and Ahnert et al [5] for determining the significant genes for the underlying biological process.

Research Method of Wilbrand et al

(1) Assign an up-down signature to a given gene or a microarray curve of N + 1 data points as follows: Connect consecutive pairs of data points by line segments and attach to each of these line segments a plus (+) if it is increasing and a minus (-) if it is decreasing. This forms a string of +'s and –'s of length N, which is its an up-down signature σ .

(2) Calculate the probability $P(\sigma)$ of an up-down signature σ as follows: Since the probability $P(\sigma)$ that N + 1 random data points have signature σ is identical to the probability that a random permutation of the integers $1,2,\cdots,N+1$ has the same signature σ, they used the formula $P(\sigma)=C(\sigma)/(N+1)!$ where $C(\sigma)$ equals the number of permutations that have the signature σ . They used the recurrence relation

$$C(i_1,\cdots i_n)=C(i_1-1,\cdots i_n)+\cdots$$
$$+C(i_1\cdots,i_n-1)$$

with boundary condition $C(\cdots,i,0,j,\cdots)=C(\cdots,i+j,\cdots)$ and $C(0,i,\cdots)=C(i,\cdots)$ where $(i,j,\cdots)$ denotes a group of i pluses, followed by a group of j minuses.

(3) Place the genes in the ascending order of the frequency $C(\sigma)$. The first gene is most likely to be correlated with the independent variable and the last gene least likely.

Research Method of Anhert et al

Ahnert et al [5] generalized the method of Wilbrand et al [21] by using concepts in the field of algorithmic information theory. Their method is as follows:

(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 permutation 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.

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

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

which is a bound on 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_{+-} which gives the number of permutations with the same pattern of rises and falls. 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. These research methods of Wilbrand et al [21] and Ahnert et al [5] have many advantages and are powerful tools to analyze large microarray data.

IV. FOULKES' METHOD & HOOK LENGTH FORMULA

In this section we will review Foulkes'[8] method for enumeration of permutations of 1,2,3,….n with a prescribed up-down sequence and the hook-length formula of Frame et al[7]. The following theorem is Theorem 2.1 of Foulkes[8].

Theorem (4.1): The number of permutations of $1, 2, \cdots n$ with a prescribed up-down sequence is $\sum g_{\vartheta\mu\eta} f(\mu)$, where

1. (η) is the partition whose Ferrers diagram has as its rim the skew-hook of the up-down sequence,
2. (ϑ) is the partition whose diagram is determined from the (η) - diagram by removal of its rim,
3. $g_{\theta\mu\eta}$ is the coefficient of the Schur-function $\{\eta\}$ in the product $\{\theta\}\{\mu\}$,
4. $f(\mu)$ is the dimension of the irreducible representation of the symmetric group S_n corresponding to (μ)

In Section V of Foulkes[8], a step by step method for calculating the coefficients $g_{\vartheta\mu\eta}$ is given. The set of all permutations of 1,2,…..,n with a given up-down (U-D) sequence is classified into disjoint subsets, each subset is characterized by the "line of route", prescribed by the standard tableau defining the subset, that is, the line joining 1,2,…..,n in succession in the tableau. As said in section 5 of Foulkes [8], this "line of route" can be regarded as a way of fitting a suitably "deformed" version of the mirror image of the skew-hook, labeled 1,2,….,n from the top, into the (μ) - diagram so that the flats (-) of the skew-hook become either horizontal lines in the route or lines going upwards from left to right, whereas the downs (+) of the skew-hook become either vertical lines in the route or lines going downwards from right to left. Further, when each node is reached in traversing the "line of route" a Ferrers diagram has been covered by the route. The line-of-route preserves the U-D sign sequence of the skew-hook.

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

Theorem (4.2): 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. MODIFICATIONS OF RESEARCH METHODS

In this section we will describe our modifications of research methods of Wilbrand et al [21] and Anhert et al [5] and explain them by examples.

Modification for research method of Wilbrand et al

1. Assign an up-down signature to a given gene or a microarray curve of N + 1 data points as described in Wilbrand et al [21].

2. Calculate the probability $P(\sigma)$ of an up-down signature σ using Foulkes' [8] method as follows: Since the probability $P(\sigma)$ that N + 1 random data points have signature σ is identical to the probability that a random permutation of the integers $1, 2, \cdots, N+1$ has the same signature σ, Wilbrand et al [21] used the formula $P(\sigma) = C(\sigma)/(N+1)!$ where $C(\sigma)$ equals the number of permutations that have the signature σ. We will calculate $C(\sigma)$ using Theorem 4.1 (which is Theorem 2.1 of Foulkes [8]) and the step-by-step method of Section V of Foulkes [8].

3. Place the genes in the ascending order of the frequency $C(\sigma)$. The first gene is most likely to be correlated with the independent variable and the last gene least likely.

Example 5.1 : Suppose the given microarray curve has the following data points: 0.41, 0.52, 0.63, 0.35, 0.21

1. By connecting these data points, the up-down signature of the microarray curve is $++--$
2. As shown in Fig. 1, if we use Theorem 4.1 (which is Theorem 2.1 of Foulkes[8]) and the step-by-step method of Foulkes [8] we get a shape of a standard Young tableau consisting of 3 elements in the first row, 1 element in the second row, and 1 element in the third row.

   So using Theorem (4.2) we get $C(\sigma) = {}^{5!}\!/_{(1.2.5.2.1)} = 6$ and $P(\sigma)$ = 6/120.

3. Place the genes in the ascending order of the frequency $C(\sigma)$. The first gene is most likely to be correlated with the independent variable and the last gene least likely.

Modification for research method of Ahnert et al

(1) Convert all microarray curves into their rank permutations as described in Ahnert et al [5].

(2) Find the values γ_{+-} of $\sigma(f)$ and $p(f)$ by using Theorem 4.1 (which is Theorem 2.1 of Foulkes[8]) and the step-by-step method of Section 5 of Foulkes[8].

(3) Compute the value $K_{\gamma}(f) = -\log_2 p(f) - \log_2 N_{\gamma}.$ for $\gamma = \gamma_{+-}$.

Example 5.2: Suppose the given microarray curve 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 and the up-down sequence is $+---$

2. As shown in Fig. 2, if we use Theorem 4.1 (which is Theorem 2.1 of Foulkes[8]) and the step-by-step method of Foulkes[8] we get a shape of a standard Young tableau consisting of 4 elements in the 1st row and 1 element in the 2nd row. So by using Theorem (4.2) we get γ_{+-} of $\sigma(f)$ = $5!\big/(1.5.3.2.1) = 4$ and $p(f)$ = 4/120.

3. Now we could compute the value

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

for $\gamma = \gamma_{+-}$. It is positive. So the gene corresponding to the curve f is significant. Thus instead of using Monte-Carlo simulation we could use Foulkes' [8] method.

VI. CONCLUSION

In the last decade, many scientists have used DNA microarray (DNA chip) technology to study many diseases caused by mutations in genes. Various computational methods are required to analyze biological data of microarray experiments and identify genes which are significant to the underlying biological processes. Wilbrand et al [21] found a new method for finding significant genes in microarray experiments. Their method is based on the probabilities of up-down signatures associated to microarray expression curves. Anhert et al [5] generalized the method of Wilbrand et al [21]. Using Monte Carlo simulation they computed various bounds on any microarray curve's algorithmic compressibility which measures its significance in the underlying biological process. The methods of Wilbrand et al [21] and Anhert et al [5] are powerful tools to analyze large microarray data.

In this paper we described modifications of research methods of Wilbrand et al [21] and Ahnert et al [5]. Using Foulkes'[8] method and hook length formula of Frame et al [7], we computed probabilities of up-down signatures of microarray curves. Additionally, we computed the bound corresponding to the map γ_{+-} (which gives the number of permutations with the same pattern of rises and falls) of Ahnert et al [5] on algorithmic compressibility of any microarray curve. Foulkes' method gives a step-by-step procedure to calculate the number of permutations with the same pattern of rises and falls. The hook length formula of Frame et al [7] gives a precise way of computing number of standard Young tableaux of a certain shape. Thus we could determine the significance of genes using combinatorial algorithms of permutations and Young tableaux. This is really wonderful!

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, "Young Tableaux 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.

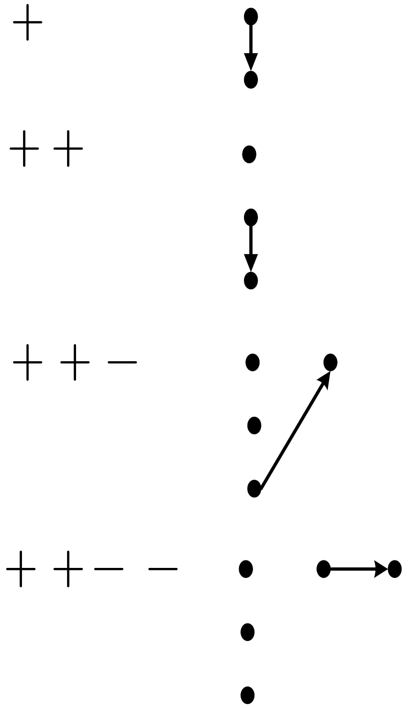

Fig. 1

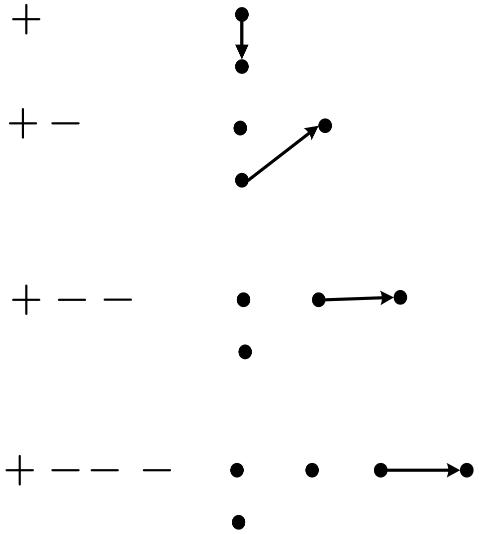

Fig. 2

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.