Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring


Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, we describe an approach to cancer classification based on gene expression monitoring by DNA microarrays. The technique involves the measurement of expression levels of thousands of genes simultaneously and the use of pattern recognition algorithms to classify tumors. We used this approach to analyze gene expression profiles of acute leukemias and non-Hodgkin's lymphomas.

The advantage of this approach is that it allows for the identification of new, previously unknown classes of tumors and the prediction of tumor class for new, previously unclassified samples. Moreover, it provides a means to identify new therapeutic targets and to monitor the response of tumors to therapy.
there were genes whose expression pattern was strongly correlated with the class distinction to be predicted. The 6817 genes were sorted by their degree of correlation (16). To establish whether the observed correlations were stronger than would be expected by chance, we developed a method called “neighborhood analysis” (Fig. 1A). Briefly, one defines an “idealized expression pattern” corresponding to a gene that is uniformly high in one class and uniformly low in the other. One tests whether there is an unusually high density of genes “nearby” (that is, similar to) this idealized pattern, as compared to equivalent random patterns.

For the 38 acute leukemia samples, neighborhood analysis showed that roughly 1100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance (Fig. 2) (17). This suggested that classification could indeed be based on expression data.

The second issue was how to use a collection of known samples to create a “class predictor” capable of assigning a new sample to one of two classes. We developed a procedure that uses a fixed subset of “informative genes” (chosen based on their correlation with the class distinction) and makes a prediction on the basis of the expression level of these genes in a new sample. Each informative gene casts a “weighted vote” for one of the classes, with the magnitude of each vote dependent on the expression level in the new sample and the degree of that gene’s correlation with the class distinction (Fig. 1B) (18, 19). The votes were summed to determine the winning class, as well as a “prediction strength” (PS), which is a measure of the margin of victory that ranges from 0 to 1 (20). The sample was assigned to the winning class if PS exceeded a predetermined threshold, and was otherwise considered uncertain. On the basis of previous analysis, we used a threshold of 0.3 (21).

The third issue was how to test the validity of class predictors. We used a two-step procedure. The accuracy of the predictors was first tested by cross-validation on the initial data set. (Briefly, one withholds a sample, builds a predictor based only on the remaining samples, and predicts the class of the withheld sample. The process is repeated for each sample, and the cumulative error rate is calculated.) One then builds a final predictor based on the initial data set and assesses its accuracy on an independent set of samples.

We applied this approach to the 38 acute leukemia samples. The set of informative genes to be used in the predictor was chosen to be the 50 genes most closely correlated with AML-ALL distinction in the known samples. The parameters of the predictor were determined by the expression levels of these 50 genes in the known samples. The predictor was then used to classify new samples, by applying it to the expression levels of these genes in the sample.

The 50-gene predictors derived in cross-validation tests assigned 36 of the 38 samples as either AML or ALL and the remaining two as uncertain (PS < 0.3) (22). All 36 predictions agreed with the patients’ clinical diagnosis.

We then created a 50-gene predictor on the basis of all 38 samples and applied it to an independent collection of 34 leukemia samples. The specimens consisted of 24 bone marrow and 10 peripheral blood samples (23). In total, the predictor made strong predictions for 29 of the 34 samples, and the accuracy was 100%. The success was notable because the collection included a much broader range of samples, including samples from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sample preparation protocols. Overall, the prediction strengths were quite high (median PS = 0.77 in cross-validation and 0.73

**Fig. 1. Schematic illustration of methodology.** (A) Neighborhood analysis. The class distinction is represented by an “idealized expression pattern” c, in which the expression level is uniformly high in class 1 and uniformly low in class 2. Each gene is represented by an expression vector, consisting of its expression level in each of the tumor samples. In the figure, the data set is composed of six AMLs and six ALLs. Gene g1 is well correlated with the class distinction, whereas g2 is poorly correlated. Neighborhood analysis involves counting the number of genes having various levels of correlation with c. The results are compared to the corresponding distribution obtained for random idealized expression patterns *c*, obtained by randomly permuting the coordinates of c. An unusually high density of genes indicates that there are many more genes correlated with the pattern than expected by chance. The precise measure of distance and other methodological details are described in (16, 17) and on our Web site (www.genome.wi.mit.edu/MPR). (B) Class predictor. The prediction of a new sample is based on “weighted votes” of a set of informative genes. Each such gene *g* votes for either AML or ALL, depending on whether its expression level *x* in the sample is closer to μAML or μALL (which denote, respectively, the mean expression levels of AML and ALL in a set of reference samples). The magnitude of the vote is w*y*, where *w* is a weighting factor that reflects how well the gene is correlated with the class distinction and *y* = (x - (μAML + μALL)/2) reflects the deviation of the expression level in the sample from the average of μAML and μALL. The weights for each class are summed to obtain total votes VAML and VALL. The sample is assigned to the class with the higher vote total, provided that the prediction strength exceeds a predetermined threshold. The prediction strength reflects the margin of victory and is defined as (Vwin - Vlose)/(Vwin + Vlose), where Vwin and Vlose are the respective vote totals for the winning and losing classes. Methodological details are described in (19, 20) and on the Web site.
in independent test) (Fig. 3A). The average prediction strength was lower for samples from one laboratory that used a very different protocol for sample preparation. This suggests that clinical implementation of such an approach should include standardization of sample preparation.

The choice to use 50 informative genes in the predictor was somewhat arbitrary. The number was well within the total number of genes strongly correlated with the class distinction (Fig. 2), seemed likely to be large enough to be robust against noise, and was small enough to be readily applied in a clinical setting. In fact, the results were insensitive to the particular choice: Predictors based on between 10 and 200 genes were all found to be 100% accurate, reflecting the strong correlation of genes with the AML-ALL distinction together with curves showing the 5 and 1% significance levels for the number of genes per cluster in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (33). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly AML (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (34). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (33). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly AML (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (33). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly AML (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (33). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly AML (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (33). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly AML (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.
that class discovery could be tested by class prediction: If putative classes reflect true structure, then a class predictor based on these classes should perform well.

To test this hypothesis, we evaluated the clusters A1 and A2. We constructed predictors to assign new samples as “type A1” or “type A2.” Predictors that used a wide range of different numbers of informative genes performed well in cross-validation. For example, a 20-gene predictor gave 34 accurate predictions with high prediction strength, one error, and three uncertains (34). The one “error” was the assignment of the sole AML sample in class A1 to class A2, and two of the three uncertains were ALL samples in class A2. The cross-validation thus not only showed high accuracy, but actually refined the SOM-defined classes: With one exception, the subset of samples accurately classified in cross-validation were those perfectly subdivided by the SOM into ALL and AML classes. The results suggest an iterative procedure for refining clusters, in which an SOM is used to initially cluster the data, a predictor is constructed, and samples not correctly predicted in cross-validation are removed. The edited data set could then be used to generate an improved predictor to be tested on an independent data set (35).

We then tested the class predictor of the A1-A2 distinction on the independent data set. In the general case of class discovery, predictors for novel classes cannot be assessed for “accuracy” on new samples, because the “right” way to classify the independent samples is not known. Instead, one can assess whether the new samples are assigned a high prediction strength. High prediction strengths indicate that the structure seen in the initial data set is also seen in the independent data set. The prediction strengths, in fact, were quite high: The median PS was 0.61, and 74% of samples were above threshold (Fig. 4B). To assess these results, we performed the same analyses with random clusters. Such clusters consistently yielded predictors with poor accuracy in cross-validation and low prediction strength on the independent data set (Fig. 4B). On the basis of such analysis (36), the A1-A2 distinction can be seen to be meaningful, rather than simply a statistical artifact of the initial data set. The results thus show that the ALL-AML distinction could have been automatically discovered and confirmed without previous biological knowledge.

We then sought to extend the class discovery by searching for finer subclasses of the leukemias. We used a SOM to divide the samples into four clusters (denoted B1 to B4). We subsequently obtained immunophenotype data on the samples and found that the four classes largely corresponded to AML, T-lineage ALL, B-lineage ALL, and B-lineage ALL, respectively (Fig. 4C). The four-cluster SOM thus divided the samples along

Fig. 3. (A) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (B) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.
another key biological distinction.

We again evaluated these classes by constructing class predictors (37). The four classes could be distinguished from one another, with the exception of B3 versus B4 (Fig 4D). The prediction tests thus confirmed the distinction corresponding to AML, B-ALL, and T-ALL, and suggested that it may be appropriate to merge classes B3 and B4, composed primarily of B-lineage ALL.

The class discovery approach thus automatically discovered the distinction between AML and ALL, as well as the distinction between B-cell and T-cell ALL. These are the most important distinctions known among acute leukemias, both in terms of underlying biology and clinical treatment. With larger sample collections, it would be possible to search for finer subclassifications. It will be interesting to see whether they correspond to existing subclassifications for AML and ALL or define new groupings perhaps based on fundamental similarities in mechanism of transformation.

In principle, the class discovery techniques above can be used to identify fundamental subtypes of any cancer. In general, such studies will require careful experimental design to avoid potential experimental artifacts—especially in the case of solid tumors. Biopsy specimens, for example, might have gross differences in the proportion of surrounding stromal cells. Blind application of class discovery could result in identifying classes reflecting the proportion of stromal contamination in the samples, rather than underlying tumor biology. Such “classes” would be real and reproducible, but would not be of biological or clinical interest. Various approaches could be used to avoid such artifacts—such as microscopic examination of tumor samples to ensure comparability, purification of tumor cells by flow sorting or laser-capture microdissection, computational analysis that excludes genes expressed in stromal cells, and confirmation of candidate marker genes by RNA in situ hybridization or immunohistochemistry to tumor sections.

Class discovery methods could also be used to search for fundamental mechanisms that cut across distinct types of cancers. For example, one might combine different cancers (for example, breast tumors and prostate tumors) into a single data set, eliminate those genes that correlate strongly with tissue type, and then cluster the samples based on the remaining genes.

We also describe techniques for class prediction, whereby samples can be automatically assigned to already-recognized classes. Creation of a new predictor involves expression analysis of thousands of genes to select a set of informative genes (we used 50 genes, although other choices also performed well) and then validating the accuracy of the assignments made on the basis of these genes. Subsequent application of the predictor then requires only monitoring the expression level of these informative genes. We described a class predictor able to accurately assign samples as AML or ALL. We have also similarly constructed a class predictor that accurately assigns ALL samples as either T-ALL or B-ALL (38). These class predictors could be adapted to a clinical setting, with appropriate steps to standardize the protocol for sample preparation. We envisage such a test supplementing rather than replacing existing leukemia diagnostics. Indeed, this would provide an opportunity to gain clinical experience with the use of expression-based class predictors in a well-studied cancer, before applying them to cancers with less well-developed diagnostics.

More generally, class predictors may be useful in a variety of settings. For example, one might combine different types of samples to discover differences that cut across distinct types of cancers. For example, one might combine different cancers (for example, breast tumors and prostate tumors) into a single data set, eliminate those genes that correlate strongly with tissue type, and then cluster the samples based on the remaining genes.

We compared six normal human kidney biopsies and six kidney tumors (renal cell carcinomas, RCCs) using the methods described for the leukemias. Neighborhood analysis showed a high density of genes correlated with the disease during the construction using 50 genes, and the predictions proved to be 100% accurate in cross-validation. The informative genes more highly expressed in normal kidney as compared to RCC included 13 metabolic enzymes, two ion channels, and three isoforms of the heavy-metal chelator metallothionein, all of which function in normal kidney physiology. Those more highly expressed included interleukin-1, an inflammatory cytokine responsible for the febrile response experienced by patients with RCC and CCND1, a D-type cyclin amplified in some cases of RCC.

13. The initial 38 samples were all derived from bone marrow aspirates performed at the time of diagnosis, before chemotherapy. After informed consent was obtained, mononuclear cells were collected by ficoll sedimentation and total RNA extracted with either Trizol (Gibco/BRL) or RNAqueous reagents (Ambion). The 27 ALL samples were derived from childhood ALL patients treated at the National Cancer Institute (DCI) protocols between 1980 and 1999. Samples were randomly selected from the leukemia cell bank based on availability. The 11 adult AML samples were similarly obtained from the National Cancer Institute (DRCI) protocols between 1980 and 1999. Samples were obtained without regard to immunophenotype, cyto genetics, or other molecular features. The independent set of leukemia samples was obtained from the DFCI childhood ALL bank (set S of samples and the set of informative genes. Parameters (a, b) are defined for each informative gene. The value a = P(gc) reflects the correlation between the expression level of g and the class distinction. The value b = [u(g) + µ(g)]/2 is a measure of the average mean log expression values in the two classes. Consider a new sample X to be predicted. Let xg denote the normalized log (expression level) of gene g in the sample (where the expression level is normalized by subtracting the mean and dividing by the SD of the expression levels in the initial set S). The vote of gene g is obtained by determining the value indicating a vote for class 1 and a negative value indicating a vote for class 2. The total vote V for class 1 is obtained by summing the absolute values of the positive votes: the sum of informative genes. While the total vote for class 2 is obtained by summing the absolute values of the negative votes.

20. The prediction strength PS is defined as PS = |Vclass| = |Vclass| (Vclass + Vclass), where Vclass and Vclass are the vote totals for the winning and losing classes. The measure PS reflects the relative margin of victory of the sample.

21. The appropriate PS threshold depends on the number of n genes in the predictor, because the PS is a sum of n variables corresponding to the individual genes, and thus its fluctuation for random input data sets varies inversely with n. See text concerning the specific choice of PS threshold.

22. In cross-validation, the entire prediction process is repeated from scratch with 37 of the 38 samples. This includes identifying the 5 informative genes to be used in the predictor and defining parameters for weighted voting.

23. The independent set of leukemia samples comprised 24 bone marrow and 22 children's cancer specimens, all obtained at the time of leukemia diagnosis. The ALL samples were obtained from the DFCI childhood ALL bank (n = 17) or St. Jude Children’s Research Hospital (SJCRH) (n = 26). A total of 38 leukemia samples were obtained in the initial data set were all derived from adult patients, the AML samples in the independent data set were derived from both adults and children. The samples were obtained from either the CALGB (adult...
Although the number of genes used had no significant effect on the outcome in this case (median PS for cross-validation ranged from 0.81 to 0.68 over a range of predictors using 10 to 200 genes, all with 0.9% error), it may matter in other instances. One approach is to vary the number of genes used, select the number that maximizes the accuracy rate in cross-validation, and then use the resulting model on the independent data set. In any case, we recommend using at least 10 genes for two reasons. Class predictors using a small number of genes may depend too heavily on any one gene and can produce spuriously high prediction strength (the “margin of victory” can occur by chance due to statistical fluctuation resulting from a small number of genes). In general, we also considered the 99% confidence line in neighborhood analysis to be the upper bound for gene selection.

Various statistical methods can be used to compare classes $C_1$, $C_2$, ..., $C_j$ (rather than Ficoll sedimentation), and RNA was prepared by an aqueous extraction (Qiagen). Thirty-three All samples were tested by cross-validation using a 50-gene predictor. Thirty-two of 33 samples were correctly assigned as T-ALL or B-ALL; the remaining sample received a PS $<0.3$, and no prediction was therefore made. Details are provided on our Web site.

41. We are grateful to S. Sallan, J. Ritz, K. Loughlin, S. Shurtleff, P. Kourlas, F. Smith, the Cancer and Leukemia Group B, and Children’s Cancer Group for providing valuable patient samples. We thank R. Klausner, D. C. Gilliland, D. Nathan, G. Daley, J. Staunton, M. Angelo, A. Leblanc, P. Lee, Z. Kilinski, G. Acton, and members of the Lander and Golub laboratories for helpful discussions. This work was supported in part by the Leukemia Society of America (T.R.G.); the National Institutes of Health and the Leukemia Clinical Research Foundation (C.D.B.); and Affymax, Millennium Pharmaceuticals, and Bristol-Myers Squibb (E.S.I.).

**Sequencing Complex Polysaccharides**

Ganesh Venkataraman, Zachary Shriver, Rahul Raman, Ram Saissekharam

Although rapid sequencing of polynucleotides and polypeptides has become commonplace, it has not been possible to rapidly sequence femto- to picomole amounts of tissue-derived complex polysaccharides. Heparin-like glycosaminoglycans (HLAGAs) were readily sequenced by a combination of matrix-assisted laser desorption ionization mass spectrometry and a notation system for representation of polysaccharide sequences. This will enable identification of sequences that are critical to HLGAG biological activities in anticoagulation, cell growth, and differentiation.

The chemical heterogeneity of polysaccharides, their structural complexity, and the lack of effective tools and methods have seriously limited the development of a sequencing approach that is rapid and practical, like that used for polynucleotides and polypeptides. This limitation is especially relevant in the study of glycosaminoglycan (GAG) complex polysaccharides, which are present at the cell surface and in the extracellular matrix (1, 2). Heparin or heparan sulfate–like glycosaminoglycans (HLAGGs), a subset of GAGs, are currently used clinically as anticoagulants, and this function of HLAGGs has been assigned to a specific pentasaccharide sequence that is responsible for binding to antithrombin III (3). Recent progress in developmental biology, genetics, and other fields has resulted in a virtual explosion in the discovery of important roles for HLAGGs in the biological activity of morphogens (4) (for example, Wingless, Decapentaplegic, and Hedgehog); growth factors, cytokines, and chemokines (5); enzymes (I, 6); and surface proteins of microorganisms (7). Although it is increasingly recognized that a specific sequence, typically from a tetra- to a decasaccharide in size, is responsible for HLAGGs’ modulation of biological activity, in only a few cases is there any structural information regarding sequences (8). Therefore, accelerating our understanding of structure-function relationships for HLAGGs requires the development of rapid yet thorough sequencing methodologies.

There are many issues that have limited the development of sequencing techniques for HLAGGs. HLAGGs are chemically complex and heterogeneous, because the HLAGG chain can vary in terms of the number of disaccharide repeat units and possesses, within the disaccharide repeat unit, four potential sites for chemical modification. The basic disaccharide repeat unit of HLAGG is a uronic acid (α-L-iduronic acid (I) or β-D-glucuronic acid (G)) linked 1,4 to an α-hexosamine (H) (Fig. 1A). Together, the four different modifications ($2^4 = 16$) for an I or G uronic acid isomer containing disaccharide give rise to $16 \times 2 = 32$ different plausible disaccharide units for HLAGGs. In contrast, four bases make up DNA, and 20 amino acids make up proteins. With these 32 building blocks, an octasaccharide could have over a million possible sequences, thereby making HLAGGs not only the most acidic but also the most information-dense biopolymers found in nature. There are no methods available to amplify or produce HLAGGs in large amounts, unlike the techniques that are available for DNA or proteins.

To handle the enormous information den-