Chemometric analysis for comparison of heparan sulphate oligosaccharides

T M Puvirajesinghe, S E Guimond, J E Turnbull, Sebastien Guenneau

JOURNAL OF THE ROYAL SOCIETY Interface

Chemometric analysis for comparison of heparan sulphate oligosaccharides

T. M. Puvirajesinghe, S. E. Guimond, J. E. Turnbull and S. Guenneau

J. R. Soc. Interface 2009 **6**, 997-1004 first published online 20 January 2009
doi: 10.1098/rsif.2008.0483

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2009) **6**, 997–1004
doi:10.1098/rsif.2008.0483
Published online 20 January 2009

Chemometric analysis for comparison of heparan sulphate oligosaccharides

T. M. Puvirajesinghe[1], S. E. Guimond[1], J. E. Turnbull[1] and S. Guenneau[2,*]

[1]*School of Biological Sciences, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK*
[2]*Department of Mathematical Sciences, University of Liverpool, Peach Street, Liverpool L69 3BX, UK*

Heparan sulphate (HS) is a glycosaminoglycan present in all metazoan organisms. It is an unbranched chain made up of repeating disaccharide units of uronic acid and glucosamine sugars, and is present in both cells and the extracellular matrix. It is one of the most structurally diverse biological molecules and its biosynthesis involves a variety of enzymic modification steps. Unlike the genome and the transcriptome, HS synthesis is not template driven. Nevertheless, the HS structure and function are highly regulated with modification steps occurring in discrete regions of the polysaccharide chain to give rise to diverse structures interacting with, and regulating, many different proteins. The resulting variation leads to diverse biological roles of HS. To study this structural diversity, rapid isolation and characterization of HS from small amounts of tissues, followed by digestion with bacterially derived enzymes (heparitinases) and chromatography techniques can be used to separate HS oligosaccharides of different size and charge. However, this leads to complex datasets where comparison of just a few samples leads to difficulties in data analysis. Using automatically integrated peak data obtained from chromatographic software, one can apply the effective disc technique to the data points to obtain the centre of mass in each dataset, for example from different murine tissues. This allows facile comparative analysis of different datasets. When the cloud of points displays some preferential direction (anisotropy), it is preferable to compute its effective ellipse. Analysis of the dynamics of the cloud of points for repeated experiments allows the quantification of their reproducibility through evaluation of an average Lyapunov exponent characterizing the area-preserving nature of a sequence of effective ellipses. These basic mathematical approaches allow a more systematic comparison of datasets derived from structural analysis using basic spreadsheet software calculations and contribute to the development of system biology strategies for tackling biocomplexity of HS polysaccharides.

Keywords: glycans; heparan sulphate; glycomics; effective shape; Lyapunov exponent

1. INTRODUCTION

Heparan sulphate (HS) is a ubiquituous linear glycosaminoglycan that is found attached to any of several families of core proteins to form complex glycoproteins. It is diverse in its cellular and extracellular distributions and is critical in a range of biological activities from development to signalling regulation (Bernfield *et al.* 1999; Turnbull *et al.* 2001). A highly sulphated version of HS, heparin, is used as a pharmaceutical drug for its anticoagulative properties. The structural diversity of HS within and between tissues is generated by the action of a complex family of biosynthetic enzymes. HS is synthesized in the Golgi on a Xyl–Gal–Gal–GlcNAc acceptor that forms a linker sequence attached to serine residues on the core protein. The sugars D-glucuronic acid and *N*-acetylglucosamine residue (GlcNAc) are added alternately to form a disaccharide repeat polymer (the nascent chain). The chain is then modified by the action of any or all of several families of enzymes. One of the first steps is the removal of the acetyl group from GlcNAc and its replacement with a sulphate group. The D-glucuronic acid can then undergo epimerization to D-iduronic acid, and either of the uronic acids can be *O*-sulphated on carbon 2. The glucosamine sugar can also be *O*-sulphated at carbon 6 or more rarely on carbon 3 (figure 1, reviewed in Esko & Lindahl (2001) and Powell *et al.* (2004)). Spatial and temporal regulations of the expression and bioactivity of the enzymes, as well as the fact that the reactions are not template driven and do not go to completion, account for the diversity in HS structure (Turnbull *et al.* 2001). These modifications occur in a regulated fashion with certain modifications occurring in discrete areas. Regions of modifications tend to cluster in highly sulphated areas (S domains) whereas stretches on unmodified disaccharides containing *N*-acetylated glucosamine residues form NA domains, as shown in figure 1 (see also Turnbull *et al.* 2001; Gallagher 2006).

*Author for correspondence (guenneau@liverpool.ac.uk).

Received 12 November 2008
Accepted 8 December 2008

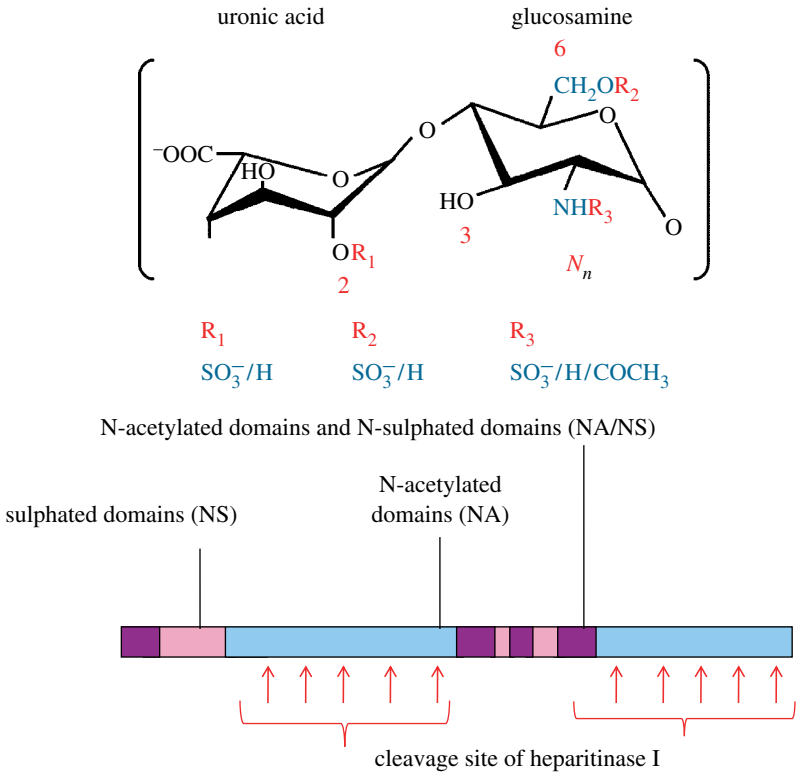

Figure 1. Structural features of HS. Disaccharide repeat unit consists of uronic acid residue and glucosamine residue. Various biosynthetic modifications can occur on the R positions of the monosaccharide units, involving functional groups, H hydroxyl, $COCH_3$ acetyl groups and SO_3 sulphate groups. R_1=H or SO_3^-, R_2=H or SO_3^-, and R_3=H_2COCH_3 or SO_3^-. Sulphated domains (S domains) contain *N*-sulphated disaccharides with IdoA-2-*O* sulphate as major uronate component. *N*-acetylated (NA) domains are non-sulphated and acetylated regions. By contrast, *N*-acetylated and *N*-sulphated (NA/NS) domains contain alternating *N*-acetylated and *N*-sulphated units. Red arrows show the cleavage sites of heparitinase I enzymes. Carbon positions are labelled with red as 2, 3 and 6 and *N*.

Biosynthesis results in HS that can differ in structure within a cell as well as between cells and tissues. HS structures are also known to vary temporally (Ford-Perriss *et al.* 2002). All these contributors result in HS having a vast set of possible structures—the 'heparanome' (Turnbull *et al.* 2001). Along with the improvements in technologies that allow high-throughput analysis, this has resulted in the emergence of methods to study the analysis of large datasets—'glycomics' (as reviewed in Turnbull & Field 2007). It has been shown that specific saccharide structures are required for activity, for example 6-*O* sulphation is essential in FGF-2 signalling using FGFR-1 (Guimond *et al.* 1993; Pye *et al.* 1998). A regulatory role of HS has also been highlighted in experimental work that has shown that selected structures are capable of activating and inhibiting FGF signalling by having specificity for different ligand and receptor isoforms (Guimond *et al.* 1993; Turnbull *et al.* 1999). In order to isolate and separate biologically active oligosaccharides, one can use commercially available enzymes that digest the HS polysaccharide in different and specific ways. This can be achieved using heparin lyases from *Flavobacterium* (Payza & Korn 1956; Lohse & Linhardt 1992), which endolytically cleave the glycosaminoglycans (GAGs). For example, in order to isolate S-domain regions of the polysaccharide, heparitinase I can be used, which cleaves polysaccharide chains containing one to four linkages between hexosamines and glucuronic acid residues, thus leaving S domains intact (as indicated in figure 1).

Analysis of S domains is achieved with strong anion exchange chromatography over a linear chloride counter ion gradient of 0–2 M NaCl allowing separation of the different structures on the basis of charge, as shown in figure 2. Once this is achieved for a series of samples, it is soon evident that comparisons become increasingly difficult owing to the complexity of the chromatographic data. Using the automatically integrated peak data obtained from chromatographic software, it is possible to get a list of all peaks with their corresponding peak heights and absorbance values for each sample, as shown in the scatter graph in figure 3. Therefore, the application of the effective disc method, which amounts to evaluating the centre of mass (or centroid) and then computing its average distance from all points within a set of data (effective radius), is a facile tool to simplify comparative analysis between spectra of this type. These calculations can be achieved using a basic spreadsheet software package, which is advantageous as other chemometric methods such as principal component analysis (PCA) have recently been used for the differentiation of polysaccharide structures from the cell wall of the tomato fruit plant (Quemener *et al.* 2007). However, although this method has similar advantage of rapidity of analysis, it does require a MATLAB-based computing environment, whereas the method of effective shape requires only basic spreadsheet software packages. Our approach is an alternative to the statistical concept of median,

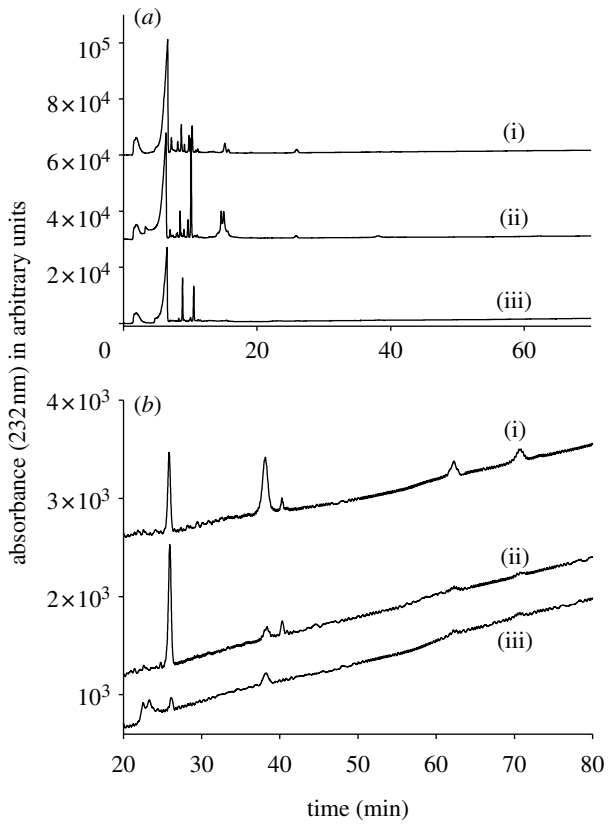

Figure 2. Strong anion exchange chromatography showing: (*a*) the separation of heparitinase I-derived HS oligosaccharides from S-domain structures of different murine organs, kidney (i), lung (ii) and heart (iii); and (*b*) the expansion of (*a*) showing detail of chromatograms from 20 to 80 min. Structures were separated on a ProPac PA1 column using a 0–2 M NaCl gradient over 90 min. Vertical axis shows the scale of (iii); other chromatograms are to the same scale.

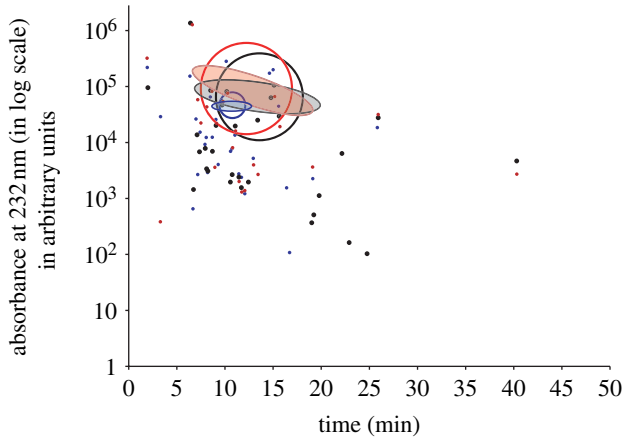

Figure 3. Data for the peak height and retention times of all the automatically integrated peaks arising from different murine tissues, heart (blue), kidney (black) and lung (red). Optimization of raw data into effective (unscaled) circular and (scaled) elliptical shapes. The area of discs provides quantitative data (see table 1 for quantitative data for ellipses).

which might become computationally demanding for large sets of data. The latter approach is commonly used in statistics and computational geometry, since it can be generalized to data in two or more dimensions.

Given a set of points, any hyperplane that goes through a centre point divides the points into two roughly equal parts: the smaller part should have at least a $1/(d+1)$ fraction of the points, where d is the space dimension. Unlike the median, a centre point needs not be one of the data points. It is usually taken to be the middle of the two middle-ordered elements of a set, call them A and B. The simplest formula for this midpoint is then just $(A+B)/2$, which is equivalent to the expression $A+(B-A)/2$. It is important to note that the former formula is more susceptible to overflow when A and B tend to be larger numbers with the same sign, which is typically the case in our study (nevertheless, there are underflow scenarios in both expressions when small signed numbers are allowed).

In this paper, we exemplify that, when there are three or more points, the midpoint can be easily and accurately evaluated through the barycentre (or centroid) and further implemented into an effective shape algorithm. Whereas such minimization problems might require in general advanced tools of calculus of variations (e.g. Lagrange multipliers for isoperimetric inequalities in Sobolev spaces; Dacorogna 2004), we describe here a very simple shape optimization method based on common sense. Effective discs and ellipses provide an alternative viewpoint on HS chromatograms, which might look fairly non-intuitive when displayed as raw data. The algorithm is easy to formulate in any space dimension and was previously applied to the analysis of wave localization (Movchan *et al.* 2007). Moreover, we propose a simple criterion for reproducibility of experiments when dealing with complex data produced in a chromatographic study, via the classical mathematical concept of average Lyapunov exponent for ergodic dynamical systems (Kingman 1968; Oseledec 1968). This criterion is universal and could be applied to many other biological applications such as mass spectrometry analysis or microarray approaches.

2. MATERIAL AND METHODS

HS was extracted and purified from different murine organs as described previously (Freeman *et al.* 2008). Tissue was taken from older normal mice being used for standard breeding programmes; mice were humanely sacrificed and all procedures conformed to UK legal requirements and institutional guidelines. HS material was solubilized in water and half the original amount of material was digested with either heparitinase I or heparitinase III enzymes. The completion of the digestion process was monitored by measuring absorbance at 232 nm at the beginning of the digest, and after 4 and 16–20 hours of incubation, and also after an overnight incubation. Heparitinase enzymes were used at 2.5 mU in 10 µl heparitinase buffer (100 mM sodium acetate, 0.1 mM calcium acetate, pH 7.0). High performance liquid chromatography (HPLC) was performed on a Propac PA1 SAX column (4.6 mm × 250 mm). Samples were injected and oligosaccharides eluted with a linear gradient of sodium chloride (0–2.0 M over 90 min). Eluant was monitored in-line for UV absorbance for unlabelled disaccharides (A_{232} nm for the detection of unsaturated non-reducing end uronate residues). Recombinant heparitinase

enzymes were purchased from IBEX technologies (Canada). Software used for the absorbance and fluorescence measurements was the SHIMADZU software: class VP chromatography data system, 4.2 S/N 971122-46*1-2 acquisition HPLC purchased from Shimadzu Deutschland Gmbl. Data analysis was achieved using MICROSOFT EXCEL 2007 software.

3. MATHEMATICAL MODELS

In the fast-growing field of chemometric techniques used for the quantitative analysis of large datasets, linear algebra plays a crucial role with the so-called factor spaces and factor-based techniques with popular techniques such as the PCA (known as PCA or factor analysis) and partial least squares (PLS) in latent variables (Kramer 1998). When chemometrians operate in a factor space, rather than the native data space, they are simply mapping their data into a new coordinate system. They are not actually changing the data itself. The operation is no more difficult than converting from Cartesian to polar coordinates (used to map, for example, electron densities or seasonal population variations). Fourier series, which is commonly used in electrical engineering to map a signal between the time domain and the frequency domain, can also be seen as a transformation of a coordinate system. Indeed, the coordinates of the signal are changed from time and amplitude to frequency and amplitude. However, the signal itself is unchanged. This is also true for the Taylor series, which is used to approximate a curve (or a surface in two dimensions or a hypersurface in higher dimensions) over a bounded region, as a series of power terms $y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$, whereby each power term x^i, $i \in \mathbb{N}$, can be considered as a new coordinate axis and each coefficient a_i is simply the new coordinate on its respective axis.

With this renewed approach, eigenvectors of a dataset in a factor space and the PCA technique (and its variation PLS) can be encompassed within the unifying concept of geometric transforms attached to the change of coordinate systems. This allows us in turn to bridge these chemometric techniques to the classical field of dynamical systems: eigenvalues computed in the principal eigenvectors (factors) for the datasets are simply the Lyapunov exponents of the associated cloud of points. The dynamics of these datasets can be therefore analysed using powerful mathematical tools developed during the past 40 years in chaos theory (see for instance the seminal papers by Kingman (1968) and Oseledec (1968)).

In this section, we adopt the former geometric viewpoint and the latter concept of average Lyapunov exponent to greatly simplify the analysis of large datasets, exemplifying our original method with a comparative analysis of HS oligosaccharides.

3.1. Effective discs

We consider a finite set of points in the real plane (a graph), $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, N\}$. We define its centroid (or barycentre) (x_c, y_c) as

$$x_c = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad y_c = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{3.1}$$

This graph can then be replaced by an effective disc centred on (x_c, y_c) and with an effective radius given by

$$R_d = \frac{1}{N}\sum_{i=1}^{N} \sqrt{\left[(x_i - x_c)^2 + (y_i - y_c)^2\right]}. \tag{3.2}$$

We note that, in a finite-dimensional vector space, all norms are equivalent. One might therefore have chosen for instance the following definition for the radius R_s:

$$R_s = \frac{1}{N}\sum_{i=1}^{N} \max[|x_i - x_c|, |y_i - y_c|], \tag{3.3}$$

which provides us with an effective square region. And of course, other norms and thus other shapes are possible, but the effective disc naturally selects points that are equally distributed (see appendix A). Importantly, the concept of effective disc can be straightforwardly generalized to higher dimensions: the definition (3.1) for the centroid R_d is generalized to compute the radius (3.2) of a sphere in three dimensions and a hypersphere otherwise, whereas (3.3) extends, respectively, to side length of cube and hypercube.

3.2. Effective ellipses

If we now want to analyse the anisotropic nature (preferential direction) of a set of N points, we need to look at an effective ellipse that makes an angle

$$\theta_c = \frac{2}{N}\sum_{i=1}^{N} \arctan\left(\frac{y_i}{x_i + \sqrt{x_i^2 + y_i^2}}\right), \tag{3.4}$$

with the horizontal x-axis.

We can then consider a local frame with axes $x'y'$ attached to the centroid (x_c, y_c), which makes an angle θ_c with the original axes xy. In this local frame, we define the effective ellipse

$$\frac{(x' - x_c)^2}{a^2} + \frac{(y' - y_c)^2}{b^2} = 1, \tag{3.5}$$

by its semi-axes

$$a = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i'^2}, \qquad b = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i'^2}. \tag{3.6}$$

We note that an ellipse can be viewed as the image of the unit circle centred at point (x_c, y_c), under a linear map associated with a symmetric matrix $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{\mathrm{T}}$, $\boldsymbol{D}$ being a diagonal matrix with the eigenvalues of $\boldsymbol{A}$, both of which are real positive, along the main diagonal, and $\boldsymbol{P}$ being a real unitary matrix having as columns the eigenvectors of $\boldsymbol{A}$.

Then the axes of the effective ellipse will lie along the eigenvectors of $\boldsymbol{A}$, and the inverse of the square root of the eigenvalues are the lengths of the semi-major and semi-minor axes a and b. In Nicolet *et al.* (2008), one can find a comprehensive derivation of the Jacobian matrices associated with maps from circles to ellipses (and to arbitrary shapes).

In our problem, the effective ellipses can be produced by multiplying the x-coordinates of all points on the effective circles by a constant, without changing the y-coordinates. This is equivalent to stretching the effective circles out in the x-direction. Note that one should multiply the resulting matrix on the left- and right-hand sides by the rotation matrix associated with the effective angle θ_c (see equation (3.4)), i.e. $\boldsymbol{A}' = \boldsymbol{R}(\theta_c)\boldsymbol{A}\boldsymbol{R}^{\mathrm{T}}(\theta_c)$, with $\boldsymbol{R}^{\mathrm{T}}(\theta_c) = \boldsymbol{R}(-\theta_c)$. The eigenvalues of $\boldsymbol{A}'$ can be computed for different samples k and their mean $(\Lambda_1 + \Lambda_2)/2$ (known as the average Lyapunov exponent in the dynamical systems literature, see Hilborn 1994) quantifies the reproducibility of the experiment. More precisely, the expression

$$\Lambda_1 + \Lambda_2 = \lim_{k \to +\infty} \frac{1}{k} \ln |\det(\boldsymbol{A}')^k| \qquad (3.7)$$

should be equal to 0 for the experiment to be reproducible, as shown in appendix B. The reproducibility criterion can be examined by measuring the centre-to-centre spacing (usually within 1%). The centre-to-centre spacing of the effective circles is very small compared with their typical size (radius). Moreover, their radii are similar (within 1% of discrepancy). Hence, we believe that there is enough evidence to claim that the experiments were indeed reproducible.

We note that, if we would start the analysis from effective squares of side length $2R_s$ instead of effective circles of diameter $2R_d$ (opting for Cartesian instead of polar coordinates), we would end up with effective rectangles instead of effective ellipses through the same mapping as before. The interpretation of the eigenvalues of the associated matrix $\boldsymbol{A}$ would repeat mutatis mutandis.

Last, these approaches of effective discs/ellipses and effective squares/rectangles can be straightforwardly extended to higher dimensions. In three dimensions, we would shape optimize the cloud of points in effective spheres/ellipsoids or effective cubes/boxes depending upon whether we use the norm (3.2) or (3.3). In higher dimensions, visualization of datasets would require consideration of, for instance, various cross sections of hyperspheres or hypercubes, hence the simplicity of our spreadsheet approach would become less obvious (use of powerful commercial softwares such as MATLAB would be unavoidable).

4. RESULTS AND DISCUSSION

Digestion of polysaccharide HS using heparitinase results in specific cleavage of HS in regions containing one to four linkages between hexosamines and glucuronic acid residues (i.e. NA domains). This results in the release of intact S domains that can then be separated and analysed using SAX-HPLC. Isolated S domains obtained from different tissues were separated using a linear 2 M salt gradient over 90 min. Standard chromatographic representation of the data is shown in figure 2. Structures are separated on the basis of charge with more highly charged structure eluting with longer column retention times. Comparison of this type of data between tissues becomes very difficult once multiple samples are considered. In order to simplify analysis, one can use the dataset of the automatically integrated peaks in order to represent each sample as a cloud of points. At this stage, we can implement the effective disc algorithm described in the previous sections to gain a viewpoint on how diverse the peak elution times are (spreading of circles), indicating how diverse the charges of the HS structures are from its centre point, and its respective location compared with other datasets. Considering our example of three different types of murine organs (as shown in figure 3), it is very simple to see at a glance that heart (blue circle) is most shifted to lower column retention times, indicating structures in the S domain are less charged than those of the lung (red circle) and kidney (black circle), which have higher column retention times, indicating more charged structures are present in the S domain. Indeed this is in keeping with the complete compositional data of murine tissues, where the kidney is known to have higher percentages of more highly charged disaccharide structures than that of the murine heart.

Table 1. Quantitative data associated with figure 3 whereby the ellipses (with their exact orientation) have been scaled up in a (large eccentricity). Note that the x variable is a time scale (minutes) and the y variable is an absorbance rate for a UV wavelength of 232 nm. Since these two variables differ by few orders of magnitude, we adopted a log scale in figure 3 to treat them on an equal footing where we needed to scale the ellipses to avoid a vanishing eccentricity in the graph.

	organ		
	heart	lung	kidney
centroid			
centroid (x_c)	11	12	14
centroid (y_c)	4.6×10^4	9.2×10^4	6.6×10^4
effective circle			
radius R_d	7.5×10^4	2.6×10^5	2.5×10^5
area (πR_d^2)	1.8×10^{10}	2.1×10^{11}	1.9×10^{11}
effective ellipse			
angle (deg.)	0	20	8.7
semi-axis a	11	1.5×10^2	1.8×10^2
semi-axis b	1.1×10^5	8.5×10^9	4.3×10^9
area (πab)	3.5×10^6	3.9×10^{12}	2.5×10^{12}

On consideration of the quantitative data presented in table 1 for the heart, lung and kidney, we are able to observe more detailed information. The barycentres of the x- and y-axes, x_c and y_c points, show that, although the area of the circle is similar in lung and kidney, there is a 2 per cent difference in total retention time in the x_c barycentres where lung has a smaller x_c value of 12 than kidney, which is 14. Note that, in the case of measuring the similarity between samples, the average difference was 0.4 per cent between the triplicate samples (0.5 min distinguishing peaks from each other, as shown in figure 4). The distinction between lung and kidney samples is aided by using elliptical shapes as this gives the direction or anisotropy of the cloud of points from each dataset. This can be considered quantitatively using the angle of rotation. In the case of heart, the angle of rotation of the ellipse shows that the ellipse follows the plane (angle of rotation is 0°). When considering kidney and lung, kidney has an 8° rotation angle whereas lung gives a rotation angle of 20°. Therefore, the lung displays a

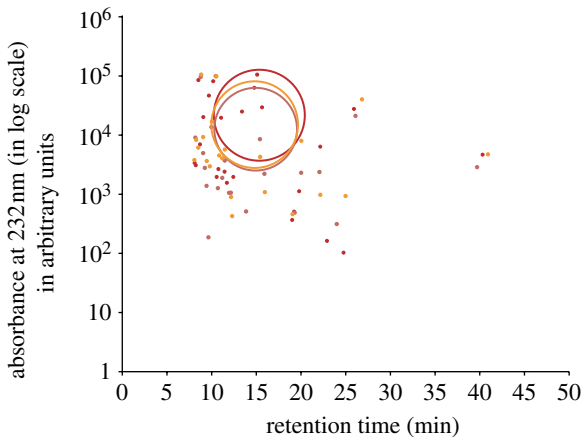

Figure 4. Three different kidney samples (yellow, red and pink) prepared on different days and separated using different preparations of the same 2 M buffer (absorbance is in milliabsorbance units). Data for automatically integrated peaks were shape optimized into circles. Reproducibility and precision of the data are visually checked. They can be measured quantitatively using the effective disc technique (by numerical comparison of centre's locations and radii, e.g. discrepancy in x_c values is 0.4% of the total 90 min and the areas of the circles are all of the same magnitude, 10^{10}). The average Lyapunov exponent describing the area preserving nature of the three corresponding effective ellipses vanishes within 1% of inaccuracy, cf. text in §3.2.

steeper gradient in the cloud of points, meaning that there is less of the population present at higher retention times than in kidney. This means that there is a smaller population of the cloud of points with more sulphated structure than in kidney. The analysis of the data shows that there is a 3 per cent difference in the angle of rotation between kidney and lung.

This technique in principle can be applied to any system that has different sets of parameters. This is most easily achieved when two parameters x and y are available. Therefore, this can be achieved for mass spectrometry data or microarray analysis, which traditionally uses complex statistical data analysis. We have shown another application of this technique in the example of separation of S-domain structures of the three murine organs separated on the basis of size using gel filtration techniques. Raw data are shown in figure 5. Data were selected by comparative elution positions with heparin standards of known size. Shape optimization of these data is shown in figure 6. Each dataset from each organ, heart (blue), kidney (black) and lung (blue), is optimized into an effective circular disc and this aids the comparison of data. Again we have used a small number of samples to exemplify the method that used the shift in the rightmost x-axis direction giving us an indication of decreasing sizes. All organs now have similar distributions in size as shown in the similar circle areas. However, when the heart structures are separated on the basis of size, the distribution is most shifted to the high retention times. Therefore, indicating that in heart the sizes of the S domains are smaller than in kidney and lung. Although again pictorially similar, the kidney and lung have a 4 per cent difference in retention times in the x_c barycentre value. Thus, the

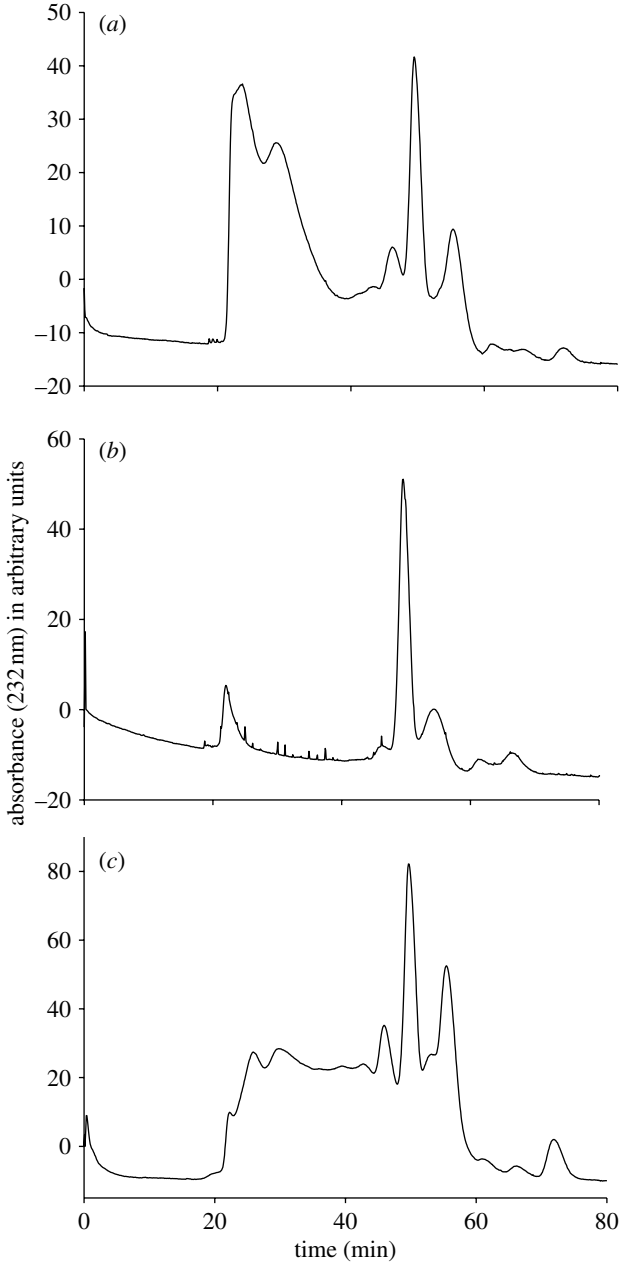

Figure 5. Gel filtration of different heparitinase I-digested products of modified TRIzol extraction products. Different murine organs were extracted and purified using the modified TRIzol extraction procedure and consequently digested using heparitinase I enzyme, therefore leaving the S-domain structures intact. The partially digested fractions were purified by gel chromatography on a Superdex peptide PE 7.5/300, run at a flow rate of 0.5 ml min^{-1} in 0.5 M ammonium hydrogen carbonate, and the elution profile was monitored by absorption at 232 nm over time (min). Peaks were identified through alignment with known authentic standards.

use of the effective circular discs allows simple patterns to be drawn out when using different experimental techniques for the same samples.

5. CONCLUSION

On consideration of the area of the effective disc and the effective ellipse (see figure 2 and table 1), the heart has the smallest area whereas the lung and kidney have a

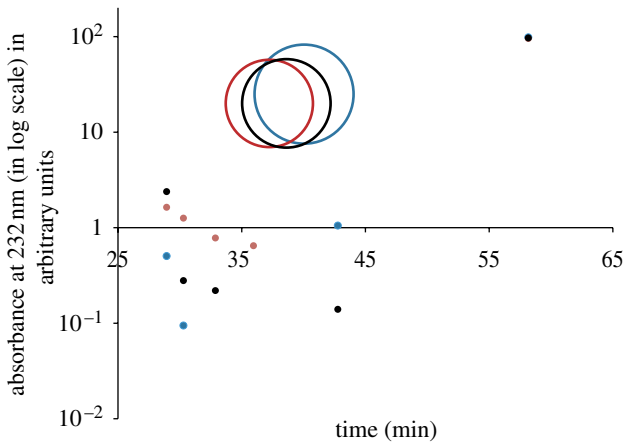

Figure 6. Optimization of HS structures separated according to size. Gel filtration raw data (figure 5) have been selected according to correspondence with heparin size standards for the murine tissues heart (blue), kidney (black) and lung (red). Therefore, the number of points of integrated peaks has been dramatically reduced. These data have then been optimized into effective circular shapes. Note that there are some overlay data points that are masked by certain spots.

much greater area by several orders of magnitude. This can then give us information about the diversity of the structures occurring in the S domain. Therefore, the information indicates that, in the murine heart, the S-domain structures are somewhat similar in their charge density to those of the kidney and lung, which have a greater diversity in their S-domain structures. We note that the areas of effective circles and ellipses differ by several orders of magnitude for all three organs: whereas the area of the blue set of points (heart) is overestimated by the effective circle (see first column of table 1), the areas of the red (lung) and black (kidney) sets of points are underestimated by the effective circle. We attribute this paradoxical tendency to the fact that murine heart is less structurally diverse than murine kidney and lung. The effective circle provides a first estimate for the effective shape of the cloud of points, whereas the effective ellipse refines this estimate and should be used for any accurate quantitative analysis.

The approach we describe can now be confidently applied to large sample sets as well as extending the approach to certain areas of interest in a chromatographic spectrum, the so-called fingerprint regions. This strategy will also be applicable to other experimental spectra such as mass spectrometry data analysis that may require rapid differentiation of different samples.

All work carried out conformed to UK legal requirement and guidelines approved by the University of Liverpool Animal Welfare Committee.

T.M.P. was supported by a PhD studentship funded by the UK Medical Research Council. S.G. is thankful for insightful comments from A. B. Movchan on the mathematical part of the paper and to S. M. Rees for pointing out references (Kingman 1968; Oseledec 1968). The authors acknowledge constructive comments from the anonymous referees, which improved the manuscript.

APPENDIX A. EQUALLY DISTRIBUTED POINTS IN THE EFFECTIVE DISC

Let us show that the points located within the effective disc can be modelled through a normal distribution, and in this way can be considered as representative of the overall cloud of points. Let us consider the points (x_i, y_i) within the effective disc, which are separated from the centre of mass (x_c, y_c) by the distance $\rho = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$. The probability that such points lie within the effective disc is

$$P(Z < R_d) = 2\pi \int_0^{R_d} f_Z(r) r \, dr, \tag{A 1}$$

where Z is the random radial variable and f_z is the Rayleigh probability density function given by (Williams 1991)

$$f_Z(\rho) = \frac{r}{\sigma_X^2} \exp(-r^2/2\sigma_X^2), \tag{A 2}$$

with σ_X the standard deviation of the random variable X, which in our case is nothing but R_d (the standard deviation of a random variable with a normal distribution is the root-mean-square deviation of its values from their mean). This shows that R_d can be interpreted as a natural measure of the statistical dispersion of points within the cloud about the centroid.

Indeed, $\sigma_X(r) = \sqrt{1/(N-1)\sum_{i=1}^{N}(r_i - r)^2}$ admits a unique minimum at $r = r_c$ (using $\sigma'(r) = 0$ at $r = r_c$ and noting that σ_X^2 is a quadratic polynomial). Hence, the standard deviation from the centroid is smaller than that from any other point within the cloud.

APPENDIX B. REPRODUCIBILITY THROUGH AVERAGE LYAPUNOV EXPONENT

To evaluate the reproducibility of a series of experiments, let us study the change of a small area A_n around a given point (x_n, y_n) within a sequence of cloud of points. This can be done through the iterative map

$$\left.\begin{aligned} x_{n+1} &= f(x_n, y_n), \\ y_{n+1} &= g(x_n, y_n), \end{aligned}\right\} \tag{B 1}$$

since the change of area can be expressed as

$$\int_{A_{n+1}} dx_{n+1} \, dy_{n+1} = \int_{A_n} |\det J| \, dx_n \, dy_n, \tag{B 2}$$

where $J = \partial(f, g)/\partial(x, y)$ is the Jacobian of (B 1) evaluated at point (x_n, y_n).

From (B 2), it follows that if:

— $|\det J| < 1$, the area A_n contracts;
— $|\det J| = 1$, the area A_n is preserved (reproducible experiments); and
— $|\det J| > 1$, the area A_n expands.

Let us now introduce two small real positive parameters ε_x and ε_y. For a chaotic discrete dynamical system, since we expect the distance between the successive images of two neighbouring points (x_1, y_1) and $(x_1 + \varepsilon_x, y_1 + \varepsilon_y)$ within the initial cloud of points to

grow exponentially with n, we write (Hilborn 1994)

$$\left.\begin{aligned}|\mathrm{d}x_n| &= |f^{(n)}(x_1+\varepsilon_x, y_1) - f^{(n)}(x_1, y_1)| \sim \varepsilon_x \exp(n\mathrm{LE}_1),\\ |\mathrm{d}y_n| &= |g^{(n)}(x_1, y_1+\varepsilon_y) - g^{(n)}(x_1, y_1)| \sim \varepsilon_y \exp(n\mathrm{LE}_2).\end{aligned}\right\} \quad \text{(B 3)}$$

Note that $\mathrm{d}x_n$ and $\mathrm{d}y_n$ correspond to the semi-axes of the sequence of effective ellipses of area $A_n = \pi \mathrm{d}x_n\, \mathrm{d}y_n$.

These two asymptotic expressions can be recast as

$$\left.\begin{aligned}\mathrm{LE}_1 &= \lim_{n\to\infty}\frac{1}{n}\left|\frac{\mathrm{d}x_n}{\varepsilon_x}\right| = \lim_{n\to\infty}\frac{1}{n}\left|\frac{\mathrm{d}x_n}{\mathrm{d}x_1}\right| = \lim_{n\to\infty}\frac{1}{n}\ln|\Lambda_1|,\\ \mathrm{LE}_2 &= \lim_{n\to\infty}\frac{1}{n}\left|\frac{\mathrm{d}y_n}{\varepsilon_y}\right| = \lim_{n\to\infty}\frac{1}{n}\left|\frac{\mathrm{d}y_n}{\mathrm{d}y_1}\right| = \lim_{n\to\infty}\frac{1}{n}\ln|\Lambda_2|,\end{aligned}\right\} \quad \text{(B 4)}$$

where Λ_1 and Λ_2 are the eigenvalues of the Jacobian

$$\boldsymbol{J} = J(x_n, y_n)J(x_{n-1}, y_{n-1})\ldots J(x_2, y_2)J(x_1, y_1). \quad \text{(B 5)}$$

LE_1 and LE_2 are the so-called Lyapunov exponents. If either (or both) $\mathrm{LE}_k > 0$, then there is a sensitive dependence on the initial condition (x_1, y_1): in other words, if $\mathrm{LE}_1 > 0$ and $\mathrm{LE}_2 < 0$ (or the way around), then there is still sensitive dependence upon (x_1, y_1). But, as it turns out, the reproducibility of experiments amounts to checking whether the average Lyapunov exponent $(\mathrm{LE}_1 + \mathrm{LE}_2)/2$ vanishes or not.

Indeed, the sum of the Lyapunov exponents satisfies

$$\begin{aligned}\mathrm{LE}_1 + \mathrm{LE}_2 &= \lim_{n\to\infty}\frac{1}{n}(\ln|\Lambda_1| + \ln|\Lambda_2|)\\ &= \lim_{n\to\infty}\frac{1}{n}\ln|\Lambda_1\Lambda_2| = \lim_{n\to\infty}\frac{1}{n}\ln|\det \boldsymbol{J}|\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln|\det J(x_i, y_i)|. \qquad \text{(B 6)}\end{aligned}$$

We conclude that if $|\det J(x_i, y_i)| = 1$ for $i = 1, \ldots, n$ (area preserving dynamical system, i.e. reproducible experiments), then $\mathrm{LE}_1 + \mathrm{LE}_2 = 0$.

REFERENCES

Bernfield, M., Gotte, M., Park, P. W., Reizes, O., Fitzgerald, M. L., Lincecum, J. & Zako, M. 1999 Functions of cell surface heparan sulfate proteoglycans. *Annu. Rev. Biochem.* **68**, 729–777. (doi:10.1146/annurev.biochem.68.1.729)

Dacorogna, B. 2004 *Introduction to the calculus of variations*. London, UK: Imperial College Press.

Esko, J. D. & Lindahl, U. 2001 Molecular diversity of heparan sulfate. *J. Clin. Invest.* **108**, 169–173. (doi:10.1172/JCI13530)

Ford-Perriss, M. *et al.* 2002 Variant heparan sulfates synthesized in developing mouse brain differentially regulate FGF signaling. *Glycobiology* **12**, 721–727. (doi:10.1093/glycob/cwf072)

Freeman, S. D., Moore, W. M., Guiral, E. C., Holme, A., Turnbull, J. E. & Pownall, E. 2008 Extracellular regulation of developmental cell signalling by XtSulf1. *Dev. Biol.* **320**, 436–445. (doi:10.1016/j.ydbio.2008.05.554)

Gallagher, J. 2006 Multiprotein signalling complexes: regional assembly on heparan sulphate. *Biochem. Soc. Trans.* **34**, 438–441. (doi:10.1042/BST0340438)

Guimond, S. E., Maccarana, M., Olwin, B. B., Lindahl, U. & Rapraeger, A. 1993 Activating and inhibitory heparin sequences for FGF-2 (basic FGF). Distinct requirements for FGF-1, FGF-2, and FGF-4. *J. Biol. Chem.* **268**, 23 906–23 914.

Hilborn, R. C. 1994 *Chaos and nonlinear dynamics*. Oxford, UK: Oxford University Press.

Kingman, J. F. C. 1968 The ergodic theory of subadditive stochastic processes. *J. R. Stat. Soc. Ser. B* **30**, 499–510.

Kramer, R. 1998 *Chemometric techniques for quantitative analysis*. New York, NY: Marcel Dekker, Inc.

Lohse, D. L. & Linhardt, R. J. 1992 Purification and characterization of heparin lyases from *Flavobacterium heparinum*. *J. Biol. Chem.* **267**, 24 347–24 355.

Movchan, A. B., Movchan, N. V., Guenneau, S. & McPhedran, R. C. 2007 Asymptotic estimates for localized electromagnetic modes in doubly periodic structures with defects. *Proc. R. Soc. A* **463**, 1045–1067. (doi:10.1098/rspa.2006.1800)

Nicolet, A., Zolla, F., Ould Agha, Y. & Guenneau, S. 2008 Geometrical transformations and equivalent materials in computational electromagnetism. *Int. J. Comput. Math. Electr. Electron. Eng.* **27**, 806–819. (doi:10.1108/03321640810878216)

Oseledec, V. I. 1968 A multiplicative ergodic theorem: liapunov characteristic numbers for dynamical systems. *Trans. Mosc. Math. Soc.* **19**, 197–231.

Payza, A. N. & Korn, E. D. 1956 Bacterial degradation of heparin. *Nature* **177**, 88–89. (doi:10.1038/177088a0)

Powell, A. K., Yates, E. A., Fernig, D. G. & Turnbull, J. E. 2004 Interactions of heparin/heparan sulfate with proteins: appraisal of structural factors and experimental approaches. *Glycobiology* **14**, 17R–30R. (doi:10.1093/glycob/cwh051)

Pye, D. A., Vives, R. R., Turnbull, J. E., Hyde, P. & Gallagher, J. 1998 Heparan sulfate oligosaccharides require 6-*O*-sulfation for promotion of basic fibroblast growth factor mitogenic activity. *J. Biol. Chem.* **273**, 22 936–22 942. (doi:10.1074/jbc.273.36.22936)

Quemener, B., Bertrand, D., Marty, I., Causse, M. & Lahaye, M. 2007 Fast data preprocessing for chromatographic fingerprints of tomato cell wall polysaccharides using chemometric methods. *J. Chromatogr. A* **1141**, 41–49. (doi:10.1016/j.chroma.2006.11.069)

Turnbull, J. E. & Field, R. A. 2007 Emerging glycomics technologies. *Nat. Chem. Biol.* **3**, 74–77. (doi:10.1038/nchembio0207-74)

Turnbull, J. E., Hopwood, J. J. & Gallagher, J. T. 1999 A strategy for rapid sequencing of heparan sulfate and heparin saccharides. *Proc. Natl Acad. Sci. USA* **96**, 2698–2703. (doi:10.1073/pnas.96.6.2698)

Turnbull, J., Powell, A. & Guimond, S. E. 2001 Heparan sulfate: decoding a dynamic multifunctional cell regulator. *Trends Cell Biol.* **11**, 75–82. (doi:10.1016/S0962-8924(00)01897-3)

Williams, R. H. 1991 *Electrical engineering probability*. New York, NY: West Publishing Company.