Characterizing temporal development of biofilm porosity using artificial neural networks

Raaja Raajan Angathevar Veluchamy, Haluk Beyenal and Zbigniew Lewandowski

ABSTRACT

We used artificial neural networks (ANN) to compute parameters characterising biofilm structure from biofilm images and to interpolate a limited number of experimental data characterising the effects of nutrient concentration and flow velocity on the areal porosity of biofilms. ANN were trained using a set of experimental data characterising structural parameters of biofilms of *Pseudomonas aeruginosa* (ATCC #700829), *Pseudomonas fluorescens* (ATCC #700830) and *Klebsiella pneumoniae* (ATCC #700831) for various flow velocities and glucose concentrations. We used 80% of the data to train ANN and 10% of the data to validate the results, which is routinely carried out as a countermeasure against overtraining. Trained ANN were used to interpolate into the data set and evaluate the missing 10% of the data. To compare ANN accuracy in evaluating the missing data with the accuracies achieved using other interpolation algorithms, we used spline, cubic, linear and nearest-neighbour interpolation algorithms to evaluate the missing data. ANN estimates were consistently closer to the experimental data than the estimates made using the other methods.

Key words | artificial neural networks, biofilms, biofilm structure

Raaja Raajan Angathevar Veluchamy
Zbigniew Lewandowski
Center for Biofilm Engineering,
Montana State University,
Bozeman, MT 59717
USA

Haluk Beyenal
School of Chemical Engineering and Bioengineering,
Washington State University,
Pullman, WA 99164
USA
E-mail: *beyenal@wsu.edu*

Zbigniew Lewandowski
Department of Civil Engineering,
Montana State University,
Bozeman, MT 59717
USA
E-mail: *zl@erc.montana.edu*

INTRODUCTION

Several research groups have demonstrated that biofilm structure controls the internal mass transfer and activity of biofilms (Wanner *et al.* 1995; Wimpenny & Colasanti 1997; Bryers & Drummond 1998; Hermanowicz 1998; Noguera *et al.* 1999*a*,*b*; Rittmann *et al.* 1999). To model the dynamics of biofilm processes it is important to quantify the relations among the biofilm structure, hydrodynamics and biofilm activity. Because biofilm activity is affected by a large number of variables, such studies are usually tedious and time-consuming. To speed up the progress of such studies, simplified methods of estimating the relations among the variables are needed. One such simplified method is interpolating into existing data sets and constructing large data sets from a limited number of measurements. Several algorithms can be used to interpolate into data sets. We tested the spline, cubic, linear, and nearest-neighbour interpolation algorithms and compared their estimates with the estimates of the artificial neural networks used as an interpolation algorithm.

The goals of this study were: (1) to verify that artificial neural networks (ANN) can be used to interpolate between existing data and to compute parameters characterising biofilm structure from biofilm images with a high degree of accuracy and (2) to use ANN to interpolate a limited number of experimental data characterising the effects of nutrient concentration and flow velocity on the areal porosity of biofilms.

To accomplish the first goal, ANN were trained in the supervised training mode, using the existing input–output pairs of experimental data, and then used to interpolate from the results of the experiments to estimate the results of the experiments that were not conducted. ANN's estimates were verified experimentally, and the accuracy of these estimates was compared with the accuracy of the estimates made using

doi: 10.2166/wst.2008.608

other available methods of interpolation, specifically the spline, cubic, linear and nearest-neighbour techniques. To accomplish the second goal, ANN were used to generate the missing data in a set of existing data quantifying the effects of glucose concentration and flow velocity on the areal porosity of biofilms. The set of experimental data composed of several parameters quantifying the temporal development of biofilm structure was generated by operating flat plate flow reactors with glass bottoms. From the biofilm images, collected through the bottom, we calculated areal porosity, fractal dimension, textural entropy, average diffusion distance, and average vertical and horizontal run lengths using the proprietary software package Image Structure Analyser-2 (ISA-2), developed at the Center for Biofilm Engineering (Yang *et al.* 2000; Beyenal *et al.* 2004). From the collected data sets, we evaluated the effects of flow velocity and glucose concentration on the parameters characterising biofilm structure. We also trained ANN to interpolate from the trends in the data sets to estimate the parameters characterising biofilm structure for several hypothetical experimental conditions. The accuracy of the estimates were tested against experimental data.

MATERIALS AND METHODS

Experimental setup

We grew the biofilms in flat plate flow reactors, as described in our previous publication (Beyenal & Lewandowski 2000). A flat plate flow reactor made of polycarbonate was 3.5 cm deep, 2.5 cm wide and 85 cm long, with a total working volume of 400 mL. The bottom of the reactor was made of glass, and the top of the reactor was sealed with silicone rubber to prevent microbial contamination. The nutrient solution was made of a pH buffer, KH_2PO_4 (0.35 g/L), Na_2HPO_4 (1.825 g/L), at pH 7.2, $(NH_4)_2SO_4$ (0.1 g/L), $MgSO_4 \times 7H_2O$ (0.01 g/L), yeast extract (0.01 g/L) and glucose, at a different concentration in each experiment. The biofilm reactors were inoculated with mixed culture of *Pseudomonas aeruginosa* (ATCC #700829), *Pseudomonas fluorescens* (ATCC #700830) and *Klebsiella pneumoniae* (ATCC # 00831).

The biofilms were grown at various combinations of average flow velocity (3.2, 10 and 25 cm/s) and glucose concentration (50, 100 and 150 mg/L). Three flow velocities and three glucose concentrations give nine pairs of flow velocity and glucose concentration at which biofilms were grown. A short hydraulic retention time, less than 20 minutes, was used to prevent the growth of microorganisms in suspension. The recycle ratio varied from 40 for the lowest flow velocity to appropriately higher ratios for the higher flow velocities. The glucose concentration in the reactor was measured using procedure 510 by Sigma® Diagnostics (St. Louis, MO). The difference in glucose concentration between the feed and the output was consistently less than 10% of the concentration in the feed. This assured us that there was no substantial difference between the rates of microbial reactions in various parts of the reactor. The nutrient solution was continuously aerated in a mixing chamber before it entered the reactor to keep the dissolved oxygen concentration near the saturation level, and the dissolved oxygen concentration was measured using a dissolved oxygen sensor.

During the operation, the reactors were attached to the stage of the Nikon® Diaphot 300 inverted microscope and the biofilm structure was monitored using incandescent light directed from the bottom. To minimise the effect of variations in light density on image quality, a constant light intensity was used (Beyenal *et al.* 2004). Images were captured by a COHU® camera (Closed Circuit, CA; model no: 2222-1040/0000) and a Flashpoint® frame grabber (Integral Technologies, Inc., Indianapolis, Indiana) connected to a computer. Each day we acquired 20 to 35 images of the biofilm through the bottom of the reactor at arbitrarily selected locations using a 40 × magnification lens. We used the Image Structure Analyser-2 (ISA-2) to calculate areal porosity, fractal dimension, textural entropy, average horizontal and vertical run lengths and average diffusion distance from the digitized biofilm images, as described in (Beyenal *et al.* 2004).

Using ANN to interpolate from the data sets to estimate the missing data

To set up ANN, we used the commercial software called EasyNN® (Neural Planner Software). The software offers a user-selectable number of hidden layers and nodes, and provides a validation of learning feature. The training process continues until the maximum normalised individual error in each cycle is less than the user-defined value.

To train ANN, we used 80% of the data sets. Ten percent of the data sets were used for validation purposes, and the remaining 10% of the data were considered missing, and were estimated by ANN. To train ANN we used the available data except those collected on the 2nd, 6th and 10th days of operation (the data sets within the shaded rectangles in Figure 1). The data not introduced into ANN, those collected on the 2nd, 6th and 10th days of operation, were considered the missing data. They represented $3 \times 3 = 9$ data sets, which is 10% of the total number of data sets. These missing data sets were then estimated using ANN, and the accuracy of these estimates was evaluated by comparing with the actual data.

During the learning process, the neural network adjusts the weight values to produce output values which closely closely the actual output values of the experimental results. This adjusting procedure is carried out by the back propagation algorithm, which calculates the root mean square (RMS) of the sum of errors (E_{RMS}) and uses this information to optimise the weight values for the next iteration. The goal of the back propagation algorithm is to minimise the E_{RMS}. To prevent overtraining, ANN was periodically validated using an additional set of data. If overtraining was detected, the number of hidden nodes in the hidden layer was lowered and a new training process was initiated.

Interpolation using spline, cubic, linear and nearest-neighbour techniques

To demonstrate that ANN can accurately estimate parameters characterising biofilm structure, we compared ANN estimates with the experimental data. We also compared the results obtained using ANN with the results obtained using other interpolation methods, specifically the spline, cubic, linear and nearest-neighbour interpolation techniques by calculating the sum of least squares (SLS) for each technique for a given set of estimates, as shown below:

$$\mathrm{SLS} = \sum_{i=1}^{N} (y_{\mathrm{predicted}} - y_{\mathrm{actual}})^2 \quad (1)$$

In the equation, N is the number of data; $y_{\mathrm{predicted}}$ is the predicted areal porosity and y_{actual} is the value measured experimentally. We used MATLAB® 7 to calculate the estimates using the spline, cubic, linear and nearest-neighbour interpolation techniques.

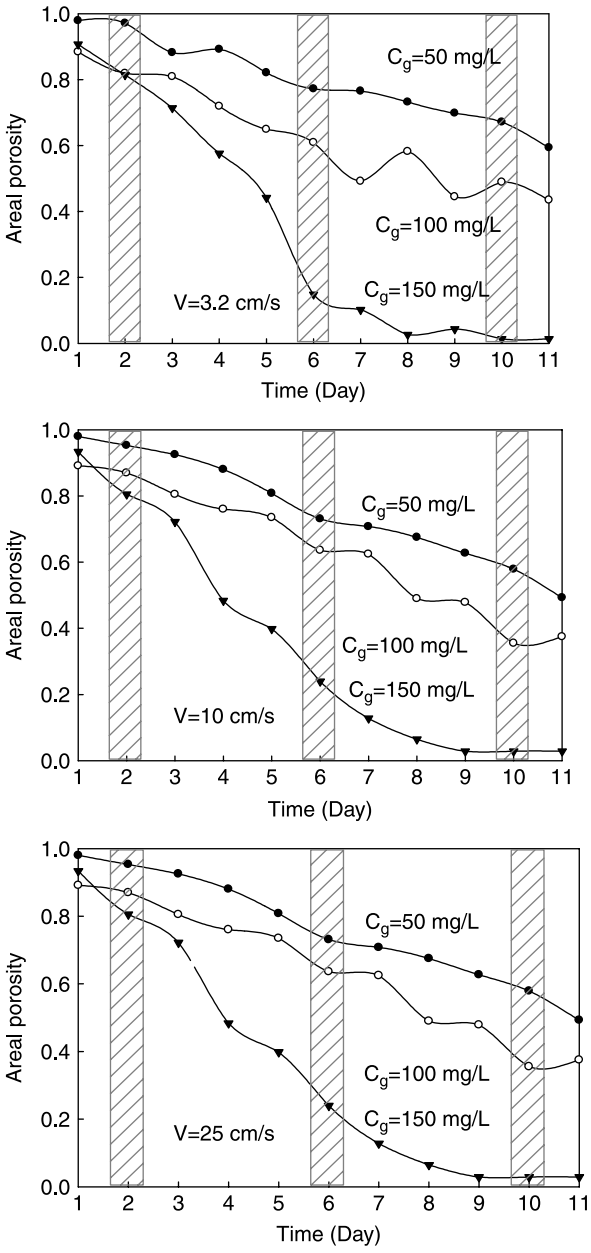

Figure 1 | Temporal variation of areal porosity in biofilms grown at various glucose concentrations (50, 100 and 150 mg/L) and various flow velocities (3.2, 10 and 25 cm/s). The data, excluding those in the shaded areas, were used to train and to validate ANN. The data in the shaded areas were used to compare ANN's estimates with the experimental data together with the estimates made using spline, cubic, linear and nearest-neighbour interpolations.

RESULTS AND DISCUSSION

Even though we calculated all the parameters specified above, for the clarity of the presentation we show only the areal porosity, which is defined as the ratio between the

combined areas of the voids and the total area of the image. Its value varies between 0 and 1; a lower areal porosity indicates higher biomass coverage at the bottom. The same procedure was used to estimate other parameters characterising biofilm structure; we have tabulated these results in Table 1.

Figure 1 shows that as time progressed, areal porosity decreased in each reactor as expected. It is also seen that areal porosities were highest in biofilms grown at low glucose concentrations and high flow velocities. These results are typical of biofilms grown for a short time before significant detachment occurs (Lewandowski *et al.* 2004). The data in Figure 1, excluding the data points in the shaded rectangular, were used to train and to validate ANN. The data in the shaded area were used to compare ANN estimates with the experimental data.

The variations of areal porosity with the glucose concentrations and flow velocities are plotted in Figure 2. The data were collected at the intersections of the grid lines only. Consequently, the data between the points where the grid lines intersect had to be evaluated by interpolation. For example, Figure 2 does not give precise information about the areal porosity of the biofilms grown at 125 mg/L glucose concentration and 15 cm/s flow velocity. To assess the areal porosities at that and other such locations we used ANN and the spline, cubic, linear and nearest-neighbour techniques.

Table 1 shows the sum of least squares (calculated from Equation 1) for each interpolation technique, and demonstrates that the complex algorithm used by ANN can successfully estimate biofilm porosity.

Figure 3 compares the estimates of areal porosity made using various interpolation techniques, and demonstrates that ANN's estimate was superior to the estimates made using other algorithms.

The remaining structural parameters we measured (Table 1) gave similar results, indicating that the estimates of missing parameters made using ANN were better than the same estimates made using the other methods (ANN had lowest sum of least squares). Having demonstrated that ANN can be used to estimate missing data by interpolating from trends in the existing sets of data, we used ANN to estimate areal porosities for various flow velocities and various substrate concentrations. The results are shown in Figure 4. To generate these results we used ANN to estimate the areal porosities in biofilms using glucose at concentrations ranging from 50 to 100 mg/L with a step size of 0.5 mg/L (a total of 200 different glucose concentrations), and grown at flow velocities ranging from 3.2 to 25 cm/s, with a step size of 0.2 cm/s (a total of 109 different flow velocities). The total number of combinations of glucose concentrations was 8,844, something practically impossible to achieve experimentally.

The purpose of the study was to demonstrate that ANN can be used as an efficient method of interpolating from trends in existing data sets, and the interpretation of the data acquired in the course of this study emphasised this particular aspect. However, with caution, the results acquired can also be interpreted from the point of view of biofilm processes. For that purpose, it is important to note that the biofilm images were acquired at the bottom level and therefore reflect the properties of the layer of the biofilm directly attached to the bottom. As expected, flow velocity and glucose concentration affected areal porosity (Figures 1 and 2). However, ascribing a specific weight to each factor associated with the variations in areal porosity is not easy: the effect does not show clear trends and it varies daily. Within the range of flow velocities and glucose concentrations applied in this study, it seems that the flow

Table 1 | The sum of least squares (calculated from Equation 1) for each interpolation technique

Method	Sum of least squares				
Parameter	Spline	Linear	Cubic	Nearest	ANN
Areal porosity	90×10^{-2}	97×10^{-2}	90×10^{-2}	2.4×10^{-2}	0.07×10^{-2}
Fractal dimension	3.9×10^{-2}	0.9×10^{-2}	3.9×10^{-2}	1.6×10^{-2}	0.1×10^{-2}
Textural entropy	0.66	0.15	0.66	0.33	0.04
Average horizontal run length	14,662	5,867	14,662	116	32
Average vertical run length	3,177	1,017	3,177	136	32
Average diffusion distance	3,482	440	3,482	1,584	40

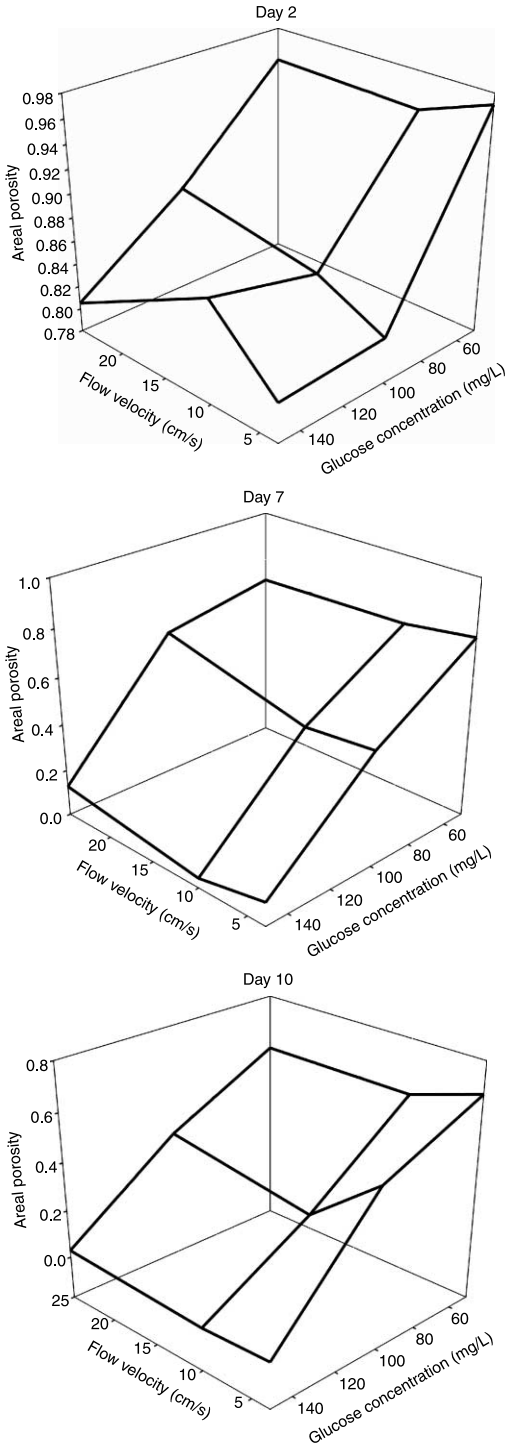

Figure 2 | Temporal variation in areal porosity versus glucose concentration and flow velocity at which the biofilms were grown for selected times. The figures are generated from the data shown in Figure 1 for the selected days of operation.

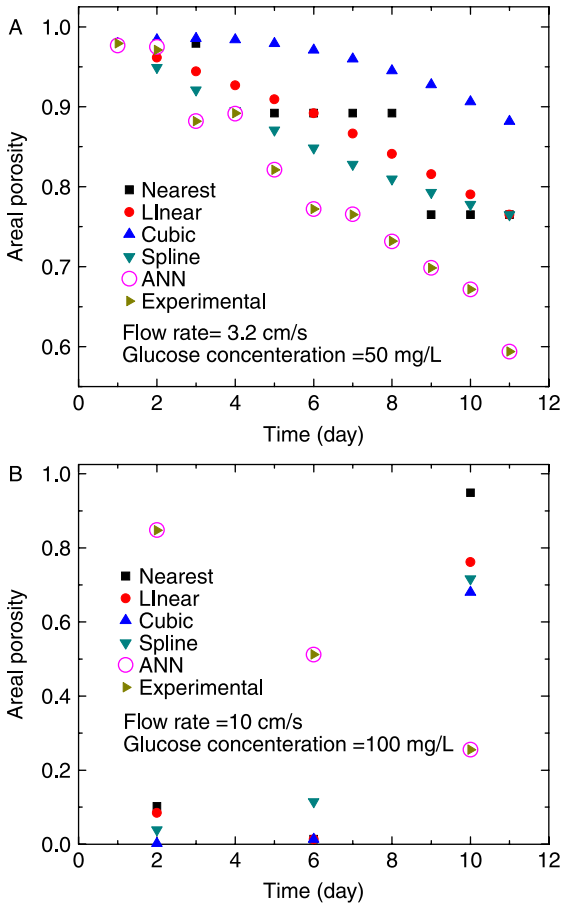

Figure 3 | The variation of areal porosity for the experimental data and estimates from tested interpolation methods. A) Estimates for the minimum flow velocity and glucose concentration at which biofilms were grown. B) Estimates for randomly selected data unseen by ANNs. In both cases ANN estimates are better than those made using the other methods tested.

velocity affected the areal porosity more than the glucose concentration did. This conclusion can be compared with a conclusion from another study in which we tested the effects of flow velocity and glucose concentration on effective diffusivity (Beyenal & Lewandowski 2000). In that study we concluded that the glucose concentration had a stronger effect on the relative effective diffusivities than the flow velocity did. Although the two studies had different goals and the ranges of glucose concentrations and flow velocities were different, we can cautiously assert that (1) glucose concentration has a stronger effect on the effective diffusivity (and density) in biofilms than flow velocity does and (2) flow velocity affects biofilm porosity to a greater extent than glucose concentration does.

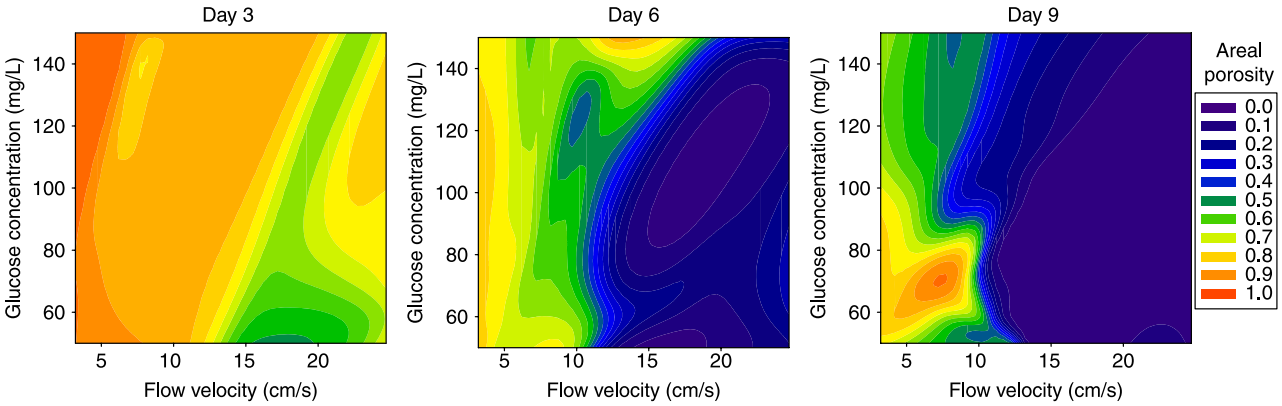

Figure 4 | Temporal variation in areal porosity estimated at varied flow velocities and substrate concentrations. The darker area indicates lower areal porosity. Subscribers to the online version of *Water Science and Technology* can access the colour version of this figure from http://www.iwaponline.com/wst.

Although we quantified the structure from the digitised images taken near the bottom of the biofilm using light microscopy, we expect that ANN can equally be used with similar images acquired using confocal scanning laser microscopy to estimate three-dimensional distributions of the parameters characterising biofilm structure.

CONCLUSIONS

ANN can be used to estimate parameters characterising biofilm structure by interpolating into an existing set of data, and these estimates are better than the estimates made using the spline, cubic, linear or nearest-neighbour interpolation techniques.

From the images of biofilms near the bottom of the reactor, we concluded that for the ranges of the variable parameters used in this study: glucose concentration between 50 and 150 mg/L and flow velocity between 3.2 and 25 cm/s, flow velocity had a stronger effect on areal porosity than glucose concentration did. The magnitude of the effects of glucose concentration and flow velocity on areal porosity varied from one day to another, which probably warrants further investigation.

ACKNOWLEDGEMENTS

Raaja Raajan Angathevar Veluchamy was supported by an award from the W. M. Keck Foundation.

REFERENCES

Beyenal, H. & Lewandowski, Z. 2000 Combined effect of substrate concentration and flow velocity on effective diffusivity in biofilms. *Water Res.* **34**, 528–538.

Beyenal, H., Lewandowski, Z. & Harkin, G. 2004 Quantifying biofilm structure: facts and fiction. *Biofouling* **20**, 1–23.

Bryers, J. D. & Drummond, F. 1998 Local macromolecule diffusion coefficients in structurally non-uniform bacterial biofilms using fluorescence recovery after photobleaching (FRAP). *Biotechnol. Bioeng.* **60**, 462–473.

Hermanowicz, S. W. 1998 A model of two-dimensional biofilm morphology. *Water Sci. Technol.* **37**(4–5), 219–222.

Lewandowski, Z., Beyenal, H. & Stookey, D. 2004 Reproducibility of biofilm processes and the meaning of steady state in biofilm reactors. *Water Sci. Technol.* **49**(11–12), 359–364.

Noguera, D. R., Okabe, S. & Picioreanu, C. 1999*a* Biofilm modeling: present status and future directions. *Water Sci. Technol.* **39**(7), 273–278.

Noguera, D. R., Pizarro, G., Stahl, D. A. & Rittmann, B. E. 1999*b* Simulation of multispecies biofilm development in three dimensions. *Water Sci. Technol.* **39**(7), 123–130.

Rittmann, B. E., Pettis, M., Reeves, H. W. & Stahl, D. A. 1999 How biofilm clusters affect substrate flux and ecological selection. *Water Sci. Technol.* **39**(7), 99–105.

Wanner, O., Cunningham, A. B. & Lundman, R. 1995 Modeling biofilm accumulation and mass-transport in a porous-medium under high substrate loading. *Biotech. Bioeng.* **47**, 703–712.

Wimpenny, J. W. T. & Colasanti, R. 1997 A unifying hypothesis for the structure of microbial biofilms based on cellular automaton models. *FEMS Microbiol. Ecol.* **22**, 1–16.

Yang, X. M., Beyenal, H., Harkin, G. & Lewandowski, Z. 2000 Quantifying biofilm structure using image analysis. *J. Microbiol. Methods* **39**, 109–119.