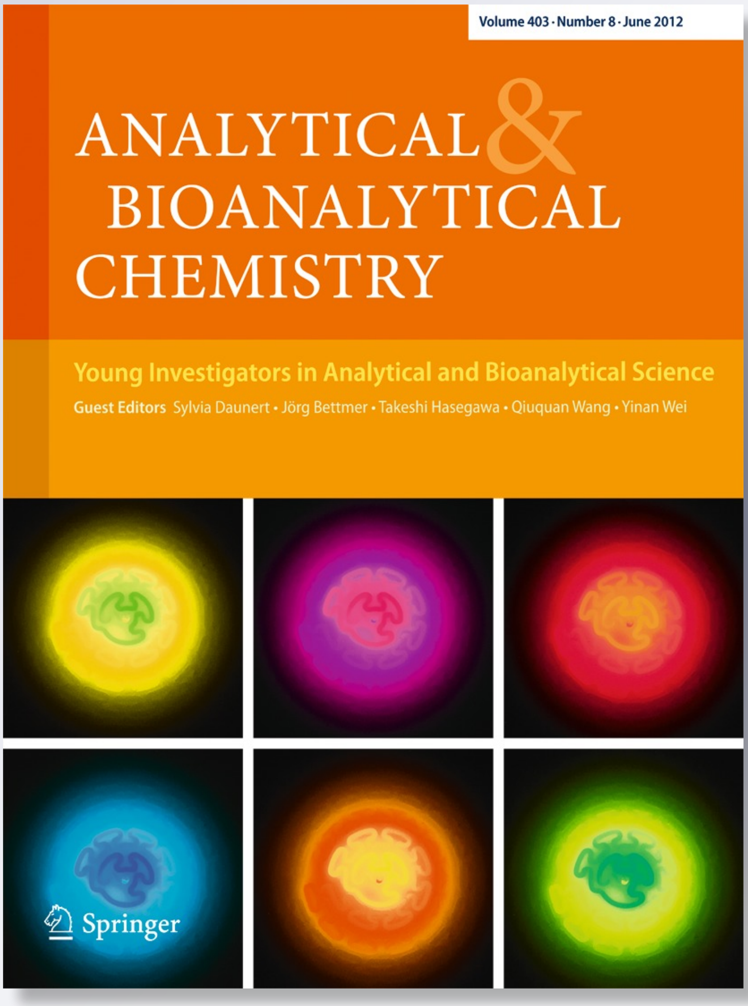
Rapid reagentless quantification of alginate biosynthesis in Pseudomonas fluorescens *bacteria mutants using FT-IR spectroscopy coupled to multivariate partial least squares regression*

Elon Correa, Håvard Sletta, David I. Ellis, Sunniva Hoel, Helga Ertesvåg, Trond E. Ellingsen, Svein Valla & Royston Goodacre

Analytical and Bioanalytical Chemistry

ISSN 1618-2642
Volume 403
Number 9

Anal Bioanal Chem (2012)
403:2591-2599
DOI 10.1007/s00216-012-6068-6



Anal Bioanal Chem (2012) 403:2591–2599
DOI 10.1007/s00216-012-6068-6

ORIGINAL PAPER

Rapid reagentless quantification of alginate biosynthesis in *Pseudomonas fluorescens* bacteria mutants using FT-IR spectroscopy coupled to multivariate partial least squares regression

Elon Correa · Håvard Sletta · David I. Ellis · Sunniva Hoel · Helga Ertesvåg · Trond E. Ellingsen · Svein Valla · Royston Goodacre

Received: 24 November 2011 / Revised: 14 March 2012 / Accepted: 19 April 2012 / Published online: 15 May 2012

Abstract Alginate is an important medical and commercial product and currently is isolated from seaweeds. Certain microorganisms also produce alginate and these polymers have the potential to replace seaweed alginates in some applications, mainly because such production will allow much better and more reproducible control of critical qualitative polymer properties. The research conducted here presents the development of a new approach to this problem by analysing a transposon insertion mutant library constructed in an alginate-producing derivative of the *Pseudomonas fluorescens* strain SBW25. The procedure is based on the non-destructive and reagent-free method of Fourier transform infrared (FT-IR) spectroscopy which is used to generate a complex biochemical infrared fingerprint of the medium after bacterial growth. First, we investigate the potential differences caused by the growth media fructose and glycerol on the bacterial phenotype and alginate synthesis in 193 selected *P. fluorescens* mutants and show that clear phenotypic differences are observed in the infrared fingerprints. In order to quantify the level of the alginate we also report the construction and interpretation of multivariate partial least squares regression models which were able to quantify alginate levels successfully with typical normalized root-mean-square error in predictions of only approximately 14 %. We have demonstrated that this high-throughput approach can be implemented in alginate screens and we believe that this FT-IR spectroscopic methodology, when combined with the most appropriate chemometrics, could easily be modified for the quantification of other valuable microbial products and play a valuable screening role for synthetic biology.

Keywords *Pseudomonas fluorescens* · Alginate · Alginate acetylation · FT-IR spectroscopy · Partial least squares regression

Introduction

Alginates are linear polysaccharides composed of mannuronic acid and guluronic acid residues produced by brown algae and bacteria such as *Azotobacter vinelandii* and some species of *Pseudomonas*. These polysaccharides are used in several large-scale industrial processes, such as food additives, and are also used in the pharmaceutical and medical sectors [1, 2]. Even though the synthesis of alginate is common among the *Pseudomonas* species, commercially

Electronic supplementary material The online version of this article (doi:10.1007/s00216-012-6068-6) contains supplementary material, which is available to authorized users.

E. Correa (✉) · D. I. Ellis · R. Goodacre
School of Chemistry, Manchester Interdisciplinary Biocentre,
University of Manchester,
131 Princess Street,
Manchester M1 7ND, UK
e-mail: elon.correa@manchester.ac.uk

H. Sletta · S. Hoel · T. E. Ellingsen
Department of Biotechnology,
SINTEF Materials and Chemistry, SINTEF,
7465 Trondheim, Norway

H. Ertesvåg · S. Valla
Department of Biotechnology, Norwegian University
of Science and Technology (NTNU),
Sem Sælandsvei 6/8,
7491 Trondheim, Norway

available alginates are currently extracted from seaweeds. The monomer composition of these alginates varies and in general one cannot control the qualitative properties of the alginates from sources that involve harvesting in nature. In contrast, most parameters can be carefully controlled in microbially produced alginates, and such products have therefore a clear potential for application—particularly in cases where product cost is not the most critical parameter. To achieve a good understanding of alginate production in relation to process development, it will be necessary to evaluate large numbers of mutants affected in alginate synthesis, and this requires highly efficient and fast screening procedures for quantification of production levels.

The research conducted here investigates the quantification of alginate biosynthesis in the non-pathogenic species *Pseudomonas fluorescens*. It has been shown that alginate production in derivatives of this organism is highly efficient [3], and transposon insertion mutants of an alginate-producing *P. fluorescens* strain were therefore used as a model to develop a new and high-throughput method for quantification of alginate. Previous work in our group has shown that FT-IR spectroscopy is an excellent vibrational spectroscopic approach [4, 5] for quantifying determinands of interest in microbial and mammalian culture including primary metabolites [6], industrially relevant secondary metabolites [7] and recombinant proteins [8, 9], but we have not yet addressed a sample that has the complexity of a very large alginate polymer.

As biological test materials in the study, we chose 193 different available transposon insertion mutants already known to be affected in alginate synthesis (mostly reduced). This set of mutants had previously been identified in a laborious screening of 10,000 mutants, each containing insertions at different and random chromosomal sites. The average alginate production value of the un-mutated control strain was used as a reference value for alginate production, such that the alginate output was then measured for each mutant strain and the reference value subtracted from the individual value observed. As a consequence, strains that produced less alginate than the control strain have a negative alginate production value assigned to them.

The mutant strains were analyzed by FT-IR spectroscopy using a high through-put scanner [10, 11]. We first report the potential differences caused by the growth medium on the bacterial phenotype and alginate synthesis in these *P. fluorescens* mutants. These results will be used to select subsequent chemometric analyses. Finally, we report the construction and interpretation of multivariate partial least squares regression (PLSR) models to quantify alginate biosynthesis successfully in *P. fluorescens* bacteria.

Methods

Cultivation of the mutant strains, enzymatic alginate quantification

For preparation of pre-cultures and cultures for the FT-IR analyses, the strains were cultivated in 96 square well plates (2 mL wells) containing 500 μL medium per well incubated at 25 °C, 800 rpm (3 mm orbital movement), 85 % humidity; 0.5× DEF4 medium (pH 7.0) was used as production media (0.35 g/L KH_2PO_4, 1,5 g/L $(NH_4)_2HPO_4$, 0,30 g/L $MgSO_4 \cdot 7H2O$, 0,49 g/L citric acid, 11.05 mg/L Fe(III) citrate, 0.57 mg/L H_3BO_3, 2,71 mg/L $MnCl_2 \cdot 2H_2O$, 2.28 mg/L $EDTA \cdot 2H_2O$, 0.28 mg/L $CuCl_2 \cdot 2H_2O$, 0.47 mg/L $Na_2MoO_4 \cdot H_2O$, 0.47 mg/L $CoCl_2 \cdot 2H_2O$, 1.40 mg/L Zn acetate$\cdot 2H_2O$, 1,1 g/L NaCl, MOPS 10 g/L, 7 g/L fructose or glycerol, and 0.5 nM m-toluate).

The pre-cultures were prepared in 0.5× DEF4 medium supplemented with 0.25 g/L yeast extract and with an increased level of C-sources (either 40 g/L fructose or glycerol). The pre-cultures were incubated overnight before they were used as inocula for the main cultures (2 vol-%) in two replicates. The main cultures with fructose and glycerol were incubated for 50 and 67 h, respectively, until the majority of the strains had utilized their carbon source. One replicate was freeze-dried and used for FT-IR analysis, the other replicate was used for enzymatic assessment of alginate yield.

The culture medium (pH 7.0) contained 7 g/L fructose or glycerol, 10 g/L MOPS 0.5 mM m-toluate. After the pre-cultures were grown overnight, 10 μL of these pre-cultures were used to inoculate the main culture medium (two replicates) and these were grown for 48 h. After 48 h, the C-source is used for the large majority of the mutants.

Alginate was quantified enzymatic essentially as described earlier in [12]. The cultures in the replicate used for enzymatic alginate measurements were supplemented with 0.05 mL/L alkalase 2.4 L and neutrase 0.5 L (Novo Nordisk) during cultivation to avoid extracellular degradation of alginate.

FT-IR data collection

All freeze-dried cultures were resuspended in 500 μL of sterile physiological saline solution (0.9 % NaCl), then 10 μL aliquots pipetted onto 96 well sample silicon sampling plates. The prepared sample plates were oven-dried for 10 min. at 50 °C to remove any excess moisture and fix the samples to the plates. The Si plates were then loaded into the motorised microplate module of a high-throughput scanner (HTS-XT™), attached to an Equinox 55 infrared spectrometer (Bruker Optics Ltd, Coventry, UK). The spectrometer was fitted with a deuterated triglycine sulfate (DTGS) detector and controlled with Opus 4, via MS Windows on an IBM compatible PC. The empty first well of each Si sample plate was used for the background measurement. Transmission spectra were collected over the wavenumber range 4,000 to 600 cm^{-1} and displayed as absorbance spectra. Collection time per spectrum was approximately 1 min. Spectra were acquired at a resolution of 4 cm^{-1}, and 64 interferograms were co-added and averaged to improve the signal-to-noise ratio. A total of three spectra were collected from a separate location of each of the samples analysed, and the infrared data were converted to ASCII format prior to statistical analysis.

To illustrate, Fig. 1 shows the FT-IR spectra of the highest and lowest alginate producing strains grouped by growth medium type. We define the highest alginate producing strain as the strain that yielded, on average, the highest quantity of alginate when compared to the reference un-mutated control strain. Likewise, we define the lowest alginate producing strain as the strain that yielded, on average, the lowest quantity of alginate when compared to that same reference un-mutated control strain (Table S1 in Electronic Supplementary Material shows the average alginate yield measured for each mutant strain studied).

Data preprocessing

Before the actual data analysis, spectra which were deviating from the natural data variability (outliers) were removed from the data set in an attempt to obtain robust models. Initially, the raw FT-IR spectra were plotted and visually examined. A small minority of the cultures used for FT-IR analysis contained very little or no biomass at all and this was reflected in the collection of poor quality spectra for those samples. For example, some spectra had exceptionally low or no absorbance at all, while others were too noisy. The objective of this data filtering phase was to generate a data set containing only reproducible and good quality spectra. We used a method based on Mahalanobis distances for multivariate spectral outlier detection. First, the method computes the Mahalanobis distance from each point to a "centre" (the mean of all data points) of the data. Second, any point with a distance to the centre larger than a predetermined cutoff value is considered an outlier [13]. Since the classical estimators of the arithmetic mean and sample covariance matrix are themselves sensitive to outliers, we used a robust estimator called minimum covariance determinant (MCD) [14] to compute the centre and the covariance matrix of the data. The cutoff value was computed as a 97.5 % quantile of the chi-square distribution with k degrees of freedom, where k is the number of variables. The algorithm used to perform this outlier detection was the "Moutlier" function from the "Chemometrics" package run in R ver. 2.9.2 (http://www.r-project.org/).

After the outliers were filtered and removed from the data, the remaining FT-IR spectra were normalized by an extended multiplicative signal correction (EMSC) algorithm. The EMSC method has originally been developed to reduce the disturbing effect of light-scattering, small particles scatter light more than larger ones [15], and we have found this method very efficient for removing unavoidable baseline shifts. This type of normalization takes the information registered in the spectra and attempts to separate physical light-scattering effects from the actual light absorbed by molecules [16].

For the computation of principal component analysis (PCA), canonical variate analysis (CVA) and partial least squares regression the data have also been autoscaled. The autoscaling process transformed columns of the data set (wavenumbers) so that each column had a mean equal to zero and a standard deviation equal to 1 [17] (see details of PCA and CVA in the Results and Discussions section).

All the computational algorithms used for the data analysis are freely available in R programming language and can

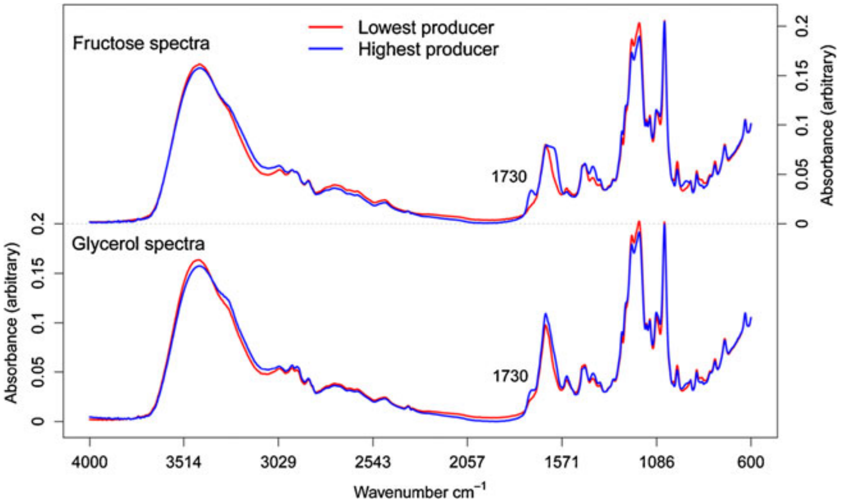


Fig. 1 FT-IR spectra from the highest (*blue*) and lowest (*red*) alginate producing strains. The plot shows spectra from fructose (*top*) and glycerol data (*bottom*)

be easily integrated to work with web applications and workflow management systems such as Taverna [18].

Results and discussion

Growth medium effect

Since the *P. fluorescens* strains were cultivated under two different growth media (fructose or glycerol), we first investigated the influence of the growth medium on the FT-IR spectra collected from these samples [19].

PCA was applied to the combined data sets, fructose + glycerol samples. The objective of PCA is to explain the variance-covariance structure of a set of variables through a few linear combinations of these variables [20]. In PCA the linear combinations (PCs) of the original variables are orthogonal (uncorrelated) to each other. Much of the original data variability can be accounted for by a small number of PCs which are then used for data reduction and visual data interpretation. For instance, given a set of data X (FT-IR spectra) with k variables (wavenumbers) PCA finds a set of vectors p_i so that $t_i = X \cdot p_i$, where $i=1, 2, \ldots, k$ and t_i (the PC scores) represents the projection coordinates of the objects (samples) from X onto the PC space and p_i (the PC loadings) represents a loadings vector. The p_i vectors are computed so that $t_1 = X \cdot p_1$ is the projection that best preserves the relative distances between the objects and t_1 is called the first principal component (PC1) which has maximum variance of the scores [21]. The explained variance then decreases from t_1 to t_2, t_2 to t_3, etc. and the k principal components reproduce the total variability of the system. The basic equation for PCA is:

$$T = X \cdot P \qquad (1)$$

where X represents the original data set, P represents the eigenvectors of the covariance matrix of X, sorted by decreasing order of eigenvalues, and T represents the PC scores or projections of the original data samples into PC space.

The results of the PCA on the combined data set are shown in Fig. 2. In this plot, two groups are observed which are related to the growth medium in which the strains were cultivated on and this indicates that the FT-IR spectra of the bacteria were dominated by these significant differences between samples grown on fructose or samples cultivated on glycerol. This suggests that the analysis of the data for the quantification of alginate should be performed individually for each growth medium type. As a result, the data were divided into fructose and glycerol data sets and all the analyses hereafter are performed and reported for each of these data sets individually.

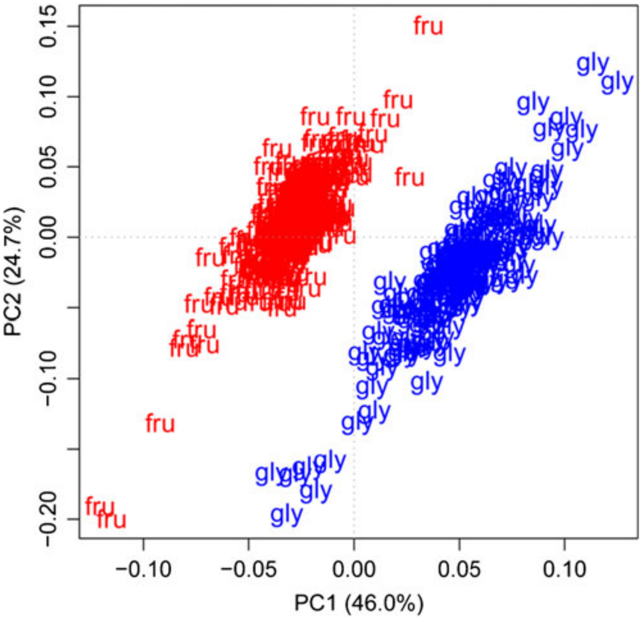


Fig. 2 Principal component analysis scores plot of the FT-IR data showing high separation of the samples by the medium used for culture growth. The total explained variance for the first two PCs is 70.7 %. The coding used was "fru" for bacteria grown on fructose as the primary carbon source and "gly" for those cultivated on glycerol

The effect of the medium on culture volumetric growth and alginate yield

As FT-IR spectroscopy detected differences between samples grown on fructose or glycerol it was important to ascertain whether this effect could be due to growth yield differences or alginate production yields, or whether this is a general phenotypic difference due to anabolism and catabolism being different, due to the organisms being grown on fructose or glycerol. The growth yield and alginate production data for all 193 mutants are shown in Table S1 in Electronic Supplementary Material.

To investigate growth yield or alginate influences, we applied two-tailed t tests to the culture growth values and alginate production measurements from the fructose and glycerol data. The objective of a two-tailed t test is to test whether the means of two sets of data (e.g., mean growth under fructose vs. mean growth under glycerol) are statistically different from each other. The null hypothesis here is that $\mu_{\text{fructose}} = \mu_{\text{glycerol}}$, where μ represents the population mean. The criterion for rejecting the null hypothesis in a two-tailed t test with 95 % confidence interval is when the total area in the two tails of the distribution is less than 5 %. Therefore, each tail must have an area of less than 2.5 % [22].

The t tests indicate that the medium type does not cause any statistically significant difference in terms of culture growth as the p value computed for this comparison was

0.1. By contrast, an alginate production effect on the bacteria was observed as the *p* value was less than 0.01. An inspection of the quantity of alginate produced by the strains shows that samples cultivated under fructose produced, on average, 62 % as much alginate as their control strain, while samples cultivated under glycerol produced only 54 % alginate relative to the control strain. On the other hand, the average quantity of alginate produced by the control strain cultivated under glycerol, 4.1 g/L, is higher than the average alginate production of the control strain cultivated under fructose 2.9 g/L.

Relationship between culture growth and alginate biosynthesis

In order to investigate the relationship between growth yield, alginate yield and FT-IR spectra further the multivariate data analysis technique of canonical variates analysis was used. In CVA given a set of predictors *X* (FT-IR spectra) and a set of grouping variables *Y* (in this case alginate production or culture growth depending on the analysis performed) the best discrimination between groups can be obtained by maximizing the Fisher ratio of the between-group variation to the within-group variation [23]. The purpose of CVA is to find linear combinations of the variables, canonical variates (CVs), which maximize this Fisher ratio. Figure 3 shows the plots of the two first sets of canonical variates computed for the FT-IR data under study collected from cells grown on either fructose (Fig. 3a, c) or glycerol (Fig. 3b, d). For each of these pairs the CVA plots the position of the objects are the same but the labels are different. Figure 3a, b shows the CVA plots using culture growth values as labels: the size of the dots depicted on the plots represent culture growth values; the larger the size of the dot is the higher the culture growth value. Figure 3c, d shows the CVA plots using alginate production values as labels. Again, the size of the dot is representative of the alginate output value. Blue dots in Fig. 3c, d represent strains that produced less alginate than the control strain.

A comparison of Fig. 3a, c indicates a strong negative correlation between culture growth yield and alginate production for the fructose data. The same relationship is observed for the glycerol data (Fig. 3b, d). Figure 3c, d also shows a grouping effect of the samples according to alginate production values. This suggests that the FT-IR analysis of the *P. fluorescens* mutant strains detected differences between the samples which are directly related to alginate biosynthesis, where when the first CV is positive the strains have a lower growth but higher alginate yield.

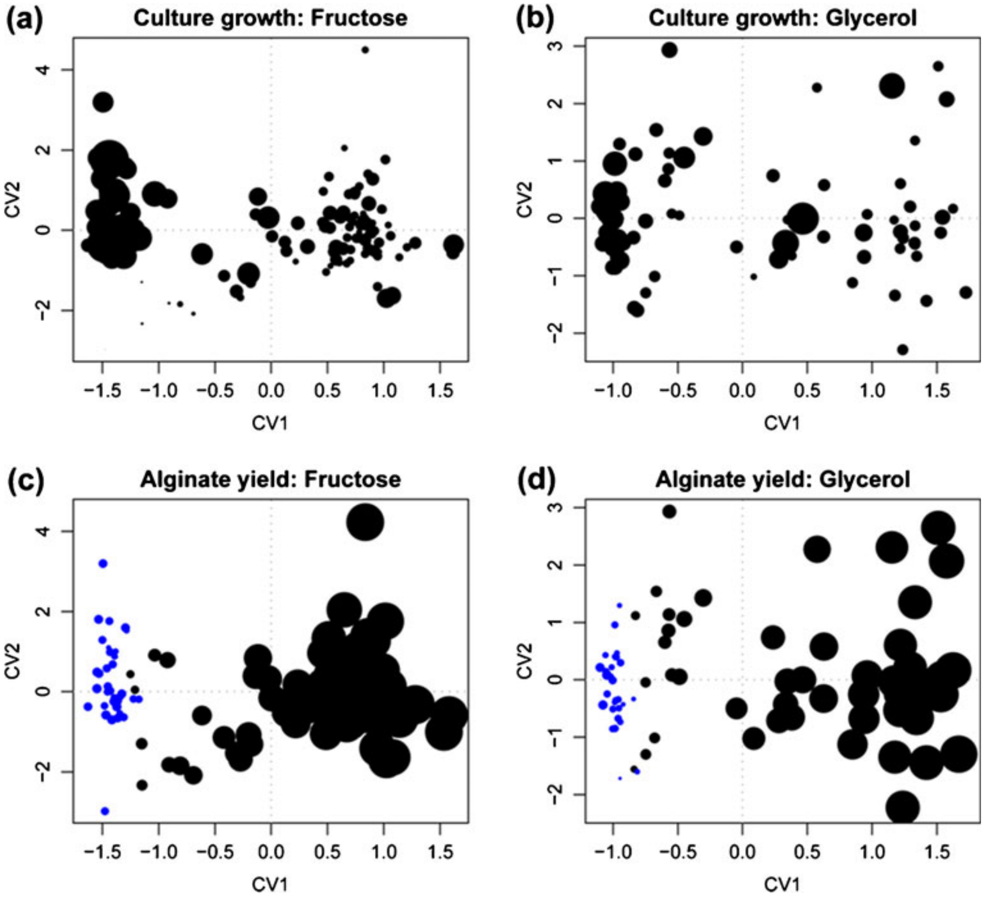


Fig. 3 CVA scores plots: culture growth measurements under fructose (**a**) and under glycerol (**b**), the size of the dots is proportional to the growth value (cellular yield). Alginate production measurements under fructose (**c**) and under glycerol (**d**), the size of the dots is proportional to the production output and blue dots represent strains that produced less alginate than the control strain

Results from multivariate PLSR applied to alginate quantification

PLSR is a supervised learning method that relates a set of independent variables X (the FT-IR data) to a set of dependent variables Y (the alginate levels). PLSR projects the X and Y variables into sets of orthogonal latent variables, scores of X and scores of Y, so that the covariance between these two sets of latent variables is maximized [24]. The purpose of PLSR is to build a linear model $Y = XB + E$, where B is a matrix of regression coefficients and E represents the difference (error) between observed and predicted Y values [25]. The size of the absolute value of the coefficient for each independent variable represents the influence of that variable on the prediction or dependent variable. The higher the absolute value of the coefficient is, the higher the influence of the variable. Once the model has been built, it can then be used to predict, or estimate, the values of the dependent variables of new samples. In addition to these predictions, loadings plots from B regression coefficients can be generated and the plot used to ascertain which variables (FT-IR absorbances) are the most important ones and will be used in model construction, and hence related to alginate (Y).

Ideally, the plot of the measured vs. the PLSR predicted values should follow the $y = x$ line or at least show a strong linear trend. Figure 4 shows the original (measured) alginate output values plotted against the ones predicted by typical PLSR models built on either the fructose (Fig. 4a) or the glycerol (Fig. 4b) data. Figure 4 shows a very good linear trend between the measured and predicted alginate values for both medium types. The normalized root-mean-square errors for the models were: fructose NRMSE=14.6 % and glycerol NRMSE=14.3 %. NRMSE is defined as the root-mean-square error (RMSE) divided by the range of observed values (Eq. 2) and displayed as percentage where low values indicate less residual variance.

$$\mathrm{NRMSE} = \left(\frac{\mathrm{RMSE}}{X_{\max} - X_{\min}}\right) \tag{2}$$

As the samples used for training also have uncertainties, these results suggest that the alginate production of *P. fluorescens* can, in fact, be quantified by PLSR. Since these PLSR models indicate successful quantification of alginate levels, we performed further tests to validate these results as follows.

Model validation is an important step in model building when using a supervised learning method. This involves the evaluation of how well the model can predict new or unseen cases not used during model building (calibration) process. Many validation methods are available for PLSR [26]. For the research conducted in this paper, we chose to use a bootstrap cross-validation method as it is a simple strategy and always provides training sets with the same number of samples as the original data set. Bootstrap is a re-sampling technique that can be applied as cross-validation to estimate the prediction performance of a model [27]. The basic idea of this method is to select randomly, with replacement, N samples from a set containing exactly N samples. All selected samples, including the repetitions, are then used as training set and the non-selected samples are used as test set [28, 29]. One can think of this as having all samples analysed ($N = \|X\|$ for our case) in a bag. A single sample is then taken out of the bag randomly and its number noted—this sample now forms part of the training data, and the sample is placed back into the bag. This random sample picking process is repeated until $\|X\|$ samples are in the training set. Some samples will be used multiple times, and on average 63.2 % of all of the samples will have been selected for training. The remainder 36.8 % are used as the test data. As we have replicate FT-IR spectra, we are careful to keep all

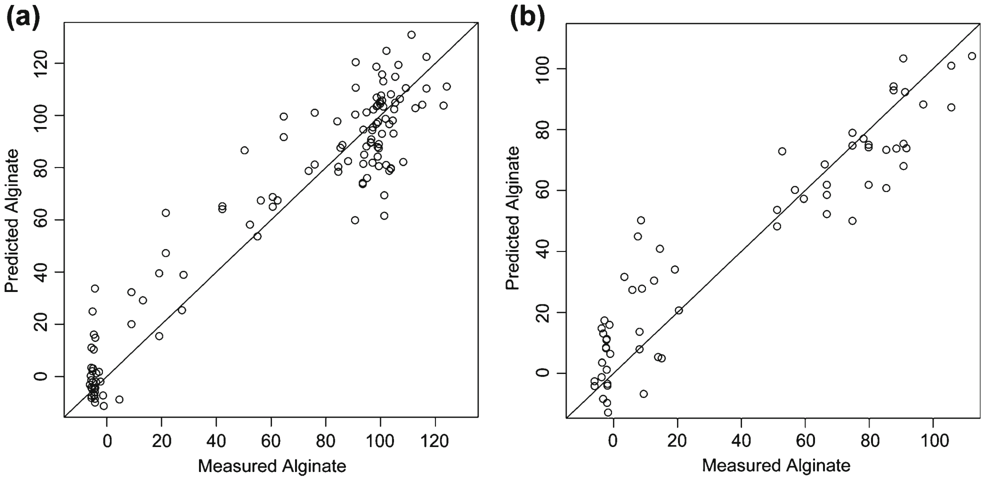


Fig. 4 Measured vs. predicted alginate values for a typical PLSR model from fructose (**a**) and glycerol (**b**)

replicate measurements together—either in the training or test sets. This process is repeated a number of times, usually 100 to 1,000 times, to try to approximate the real distribution of samples in the global population of cases. The reasoning behind bootstrap is that if the set of samples available represents the global population of cases, then re-sampling from this set is equivalent to re-sampling from the global population from which the samples were drawn.

To validate the PLSR models and to confirm that the predictions are not occurring "just by chance" (that is to say, the results are not too optimistic), we also applied a set of permutation tests to the bootstrap cross-validated models. In a permutation test the original class labels, or values, are randomly swapped [30–32] and this allows the generation of a null or random model. A prediction model is then built on these permutated data and the model is evaluated, and this process is repeated several times. The mean predictive accuracy value is computed over all permutated models and is compared to the mean predictive accuracy value computed over the models that used the original class values (alginate levels). If the predictive accuracy of the original models is significantly higher than the one computed over the permutation test models, then the evidence suggests that the effects observed in the original data are also present in the global population and the original models are indeed valid.

We applied the bootstrap PLSR model, with and without permutation testing, to the FT-IR data to measure alginate production. For the PLSR model, the X data are the spectra (FT-IR absorbance values) and the Y data are the alginate yield measured for each sample. The PLSR models were built as follows. First the data were divided into fructose and glycerol samples. A prediction model was then built based only on the fructose samples. The process is repeated 1,000 times and the results were recorded. The same procedure was then applied for the glycerol data. Figure 5 shows the cross-validated R^2 values computed for each data set on the original data and on the permutation test data. A cross-validated R^2 value close to or lower than 0 indicates that the model cannot predict the alginate outcome well whereas a value of 1 indicates that the model can perfectly predict the alginate outcome with 100 % precision. In practice, however, a R^2 value equal to 1 is almost never found and the model is considered efficient if its R^2 value is a value close to 1 such as >0.7 [25]. The cross-validated R^2 is computed after each bootstrap cross-validation according to Eq. 3:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \tag{3}$$

where $\hat{y}_i$ is the predicted value for the ith sample from the test data, y_i its observed value, $\bar{y}_i$ is the average of the observed values and n is the total number of data samples in the test data.

The bootstrap cross-validation results presented in Fig. 5a suggest that the PLSR models did detect a relationship between the *P. fluorescens* samples and their respective alginate production levels, since that the median of both R^2 values is higher than 0.7. Figure 5a indicates that the median R^2 value is higher for the fructose models ($R^2 \approx 0.86$) than it is for the glycerol models (median $R^2 \approx 0.76$).

In order to confirm that the models are actually predicting alginate outcome, we applied a set of permutation tests to the data. Any model built on these data should, obviously, give a poor prediction of alginate outcome. The results for the PLSR using permutation test are shown in Fig. 5b. The permutation test results indicate a poor prediction of alginate outcome as expected; median $R^2 < 0$ for both medium types. This agrees with the theory and, as a result, this implies that the PLSR bootstrap models built on the original data are, in fact, correctly predicting or quantifying alginate production levels of the *P. fluorescens*.

As discussed in the Materials section PLSR also generates loadings plots and further examination of these PLSR

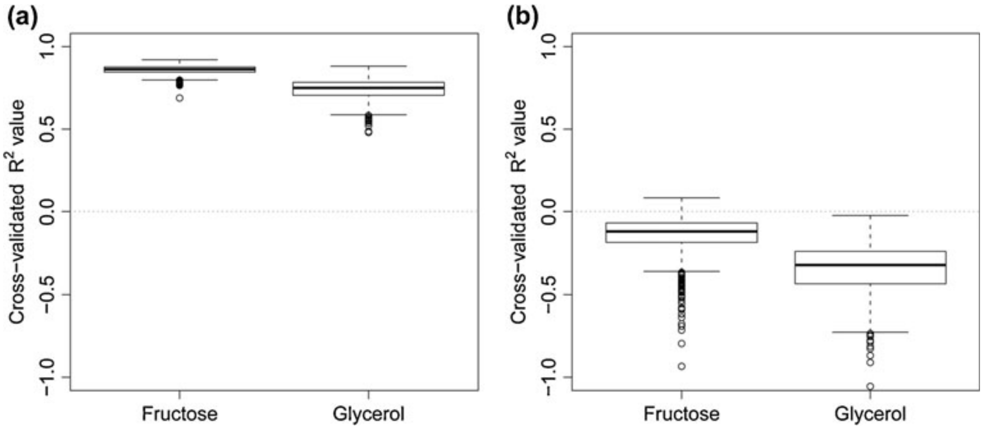


Fig. 5 Box-and-whisker plots: results from 1,000 independent bootstrap cross-validation PLSR models without permutation test (**a**) and with permutation test (**b**). A box-and-whisker plot graphically displays numeric data through their 5 number summaries: the smallest observation, lower quartile, median, upper quartile, and largest observation

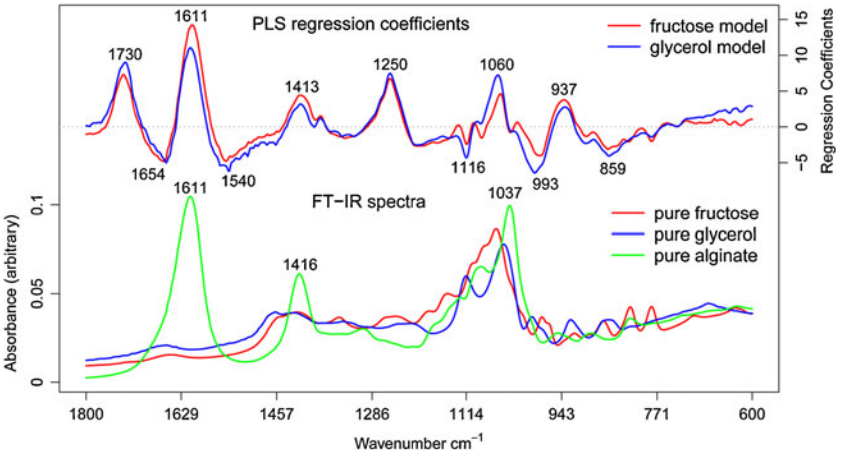


Fig. 6 PLS regression coefficients (*top*) and FT-IR spectra of pure fructose, pure glycerol, and pure alginate (*bottom*)

regression coefficients also confirms that the models built on the FT-IR data did detect alginate directly, rather than some spuriously correlated changes in the bacterial phenotype. Figure 6 shows the FT-IR spectra of pure fructose, pure glycerol, and pure alginate and the PLSR regression coefficients from a typical model. The PLSR regression coefficients were built based on the 10 first latent variables, which explain 99.34 % of the variability on the fructose data and 99.62 % on the glycerol data. A comparison of the alginate spectrum (green line) and the PLSR regression coefficients shown in Fig. 6 show that the most significant coefficient values are around 1,611 cm^{-1}, which is a characteristic peak in the spectra of pure alginate analysed by FT-IR in this study and also reported elsewhere [33]. This confirms that the PLSR models are indeed detecting alginate production in these *P. fluorescens* mutants.

Additional confirmation of this comes from a comparison between the second highest set of regression coefficients, around 1,730 cm^{-1}, and the spectra from the highest and lowest alginate producing strains shown in Fig. 1. The spectra depicted in Fig. 1 also show a characteristic difference between the lowest and the highest alginate producing strains at 1,730 cm^{-1}, although it is a weak absorption band. This band can be assigned to C=O carbonyl stretching [34–36], which has previously been observed in alginate [37, 38] and has been related to alginate acetylation [39, 40]. Thus, this reinforces the argument that PLSR is indeed detecting alginate production directly. In general, alginate acetylation seems to show two characteristic FT-IR absorption peaks: one at 1,730 cm^{-1} and another at 1,250 cm^{-1} [41]. Examination of the regression coefficients shown in Fig. 6 positively confirms the presence of such peaks. In addition, a correlation analysis of the area under the band between 1,700–1,760 cm^{-1} and alginate production shows a Pearson's correlation coefficient of 0.84 with a p value much smaller than 0.01 (p value=2.5×10^{-97}). This correlation is high, statistically significant and confirms that the PLSR model is clearly detecting the chemical reaction of acetylation on the alginate produced by the *P. fluorescens* samples.

Conclusions

The results from the analyses performed demonstrates the successful application of a practical, effective and informative multivariate regression model to the quantification of alginate from several *P. fluorescens* strains interrogated by FT-IR spectroscopy. The regression model has shown that FT-IR analysis detected biochemical variations among the strains that are directly associated to alginate production and allowed the model to correctly quantify those levels of alginate production.

In addition, FT-IR analysis was also able to detect the chemical reaction of acetylation, a reaction that may naturally occur in the alginate produced by *Pseudomonas* as demonstrated in the literature. Therefore, FT-IR spectroscopy coupled with PLSR analysis can also be used as a fast method to measure alginate acetylation.

Furthermore, analysis for the rapid, accurate, and high-throughput identification of exactly which *P. fluorescens* strains are more susceptible to the acetylation of alginate, and its possible effects on alginate production, would also be an interesting area for future studies—as would modifying the growth environment of the bacteria for enhancing alginate (or indeed any other product) yield. In order to achieve this, FT-IR spectroscopy would remain the same but as the phenotype of the bacteria will change and FT-IR generates a whole organism or whole medium fingerprint, the chemometric models would have to be recalibrated with exemplar data so that they include this new biology or phenotype.

Acknowledgements E.C., H.E., U.W., S.V. and R.G. would like to thank the "Systems Biology of Microorganisms" (SysMO) project and E.C. and R.G. in particular the Biotechnology and Biological Sciences Research Council (BBSRC) for their financial support.

References

1. Uludag H, De Vos P, Tresco PA (2000) Technology of mammalian cell encapsulation. Adv Drug Deliv Rev 42:29–64
2. Skaugrud O, Hagen A, Borgersen B, Dornish M (1999) Biomedical and pharmaceutical applications of alginate and chitosan. Biotechnol Genet Eng Rev 16:23–40
3. Gimmestad M, Sletta H, Karunakaran P, Bakkevig K, Ertesvåg H, Ellingsen T, Skjåk-Bræk G,Valla S (2004) New mutant strains of *Pseudomonas fluorescens* and variants thereof, methods for their production, and uses thereof in alginate production. (PCT) WO 2004/011628 and U.S. Patent 7553656 B2
4. Ellis DI, Goodacre R (2006) Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst 131:875–885
5. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R (2007) Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics 8:1243–1266
6. Winson MK, Goodacre R, Woodward AM, Timmins ÉM, Jones A, Alsberg BK, Rowland JJ, Kell DB (1997) Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. Anal Chim Acta 348:273–282
7. McGovern AC, Broadhurst D, Taylor J, Kaderbhai N, Winson MK, Small DA, Rowland JJ, Kell DB, Goodacre R (2002) Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. Biotechnol Bioeng 78:527–538
8. McGovern AC, Ernill R, Kara BV, Kell DB, Goodacre R (1999) Rapid analysis of the expression of heterologous proteins in Escherichia coli using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to a2-interferon production. J Biotechnol 72:157–167
9. Sellick CA, Hansen R, Jarvis RM, Maqsood AR, Stephens GM, Dickson AJ, Goodacre R (2010) Rapid monitoring of recombinant antibody production by mammalian cell cultures using Fourier transform infrared spectroscopy and chemometrics. Biotechnol Bioeng 106:432–442
10. Winder CL, Cornmell R, Schuler S, Jarvis RM, Stephens GM, Goodacre R (2011) Metabolic fingerprinting as a tool to monitor whole cell biotransformations. Anal Bioanal Chem 399: 387–401
11. Harrigan GG, LaPlante RH, Cosma GN, Cockerell G, Goodacre R, Maddox JF, Ludyenck JP, Ganey PE, Roth RA (2004) Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. Toxicol Lett 146:197–205
12. Ostgaard K (1993) Determination of alginate composition by a simple enzymatic assay. Hydrobiologia 261:513–520
13. Hardin J, Rocke DM (2004) Outlier detection in the multiple clusters setting using the minimum covariance determinant estimator. Comput Stat Data Anal 44:625–638
14. Rousseeuw P (1985) Multivariate estimation with high breakdown point. In mathematical statistics and applications, Volume B. Dordrecht-Reidel, Netherlands, pp 283–297
15. Næs T, Isaksson T, Kowalski B (1990) Locally weighted regression and scatter correction for near-infrared reflectance data. Anal Chem 62:664–673
16. Martens H, Nielsen JP, Engelsen SB (2003) Light scattering and light absorbance separated by extended multiplicative signal correction. Application to Near-Infrared transmission analysis of powder mixture. Anal Chem 75:394–404
17. Goodacre R, Broadhurst D, Smilde AK, Kristal BS et al (2007) Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3(3):231–241
18. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock M, Li P, Oinn T (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34:729–732
19. Kaderbhai NN, Broadhurst DI, Ellis DI, Goodacre R, Kell DB (2003) Functional genomics via metabolic footprinting: monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. Comp Funct Genomics 4:376–391
20. Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. Prentice Hall, New Jersey
21. Raykov T, Marcoulides GA (2008) An introduction to applied multivariate analysis. Routledge Academic, New York
22. Ennos R (2006) Statistical and data handling skills in biology. Prentice Hall, New Jersey
23. Hair JF, Black B, Babin B, Anderson RE, Tatham RL (2005) Multivariate data analysis, 6th edn. Pearson Education, New Jersey
24. Martens H, Næs T (1992) Multivariate calibration. Wiley, New York
25. Vinzi EV, Chin WW, Henseler J, Wang H (eds) (2010) Handbook of partial least squares. Springer, New York
26. Barker M (2010) Partial least squares for discrimination: Statistical theory and implementation. Lap Lambert Academic Publishing, Saarbrücken
27. Varmuza K, Filzmoser P (2009) Introduction to multivariate statistical analysis in chemometrics. CRC, Florida
28. Efron B (1981) Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika 68(3):589–599
29. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman and Hall/CRC, New York
30. Welch WJ (1990) Construction of permutation tests. J Am Stat Assoc 85:693–698
31. Good P (2002) Extensions of the concept of exchangeability and their applications. J Mod Appl Stat Methods 1:243–247
32. Golland P, Liang F, Mukherjee S, Panchenko D (2005) Permutation tests for classification. Lect Notes Comput Sci 3559:501–515
33. Sarmento B, Martins S, Ribeiro A, Veiga F, Neufeld R, Ferreira D (2006) Development and comparison of different nanoparticulate polyelectrolyte complexes as insulin carriers. Int J Pept Res Ther 12:131–138
34. Lin SY, Cheng WT, Wei WS, Lin HL (2011) DSC-FTIR microspectroscopy used to investigate the heat-induced intramolecular cyclic anhydride formation between Eudragit E and PVA copolymer. Polym J 43:577–580
35. Sarkisova S, Patrauchan MA, Berglund D, Nivens DE, Franklin MJ (2005) Calcium-induced virulence factors associated with the extracellular matrix of mucoid *Pseudomonas aeruginosa* biofilms. J Bacteriol 187:4327–4337
36. Morris NM, Catalano EA, Kottes BA (1995) FT-IR Determination of degree of esterification in polycarboxylic acid cross-link finishing of cotton. Cellulose 2:31–39
37. Nivens DE, Dennis EO, Williams J, Franklin MJ (2001) Role of alginate and its O acetylation in formation of *Pseudomonas aeruginosa* microcolonies and biofilms. J Bacteriol 183:1047–1057
38. Zhu H, Ji J, Lin R, Gao C, Feng L, Shen J (2002) Surface engineering of poly(DL-lactic acid) by entrapment of alginate-amino acid derivatives for promotion of chondrogenesis. Biomaterials 23:3141–3148
39. Franklin MJ, Dennis EO (1993) Identification of *algF* in the alginate biosynthetic gene cluster of *Pseudomonas aeruginosa* which is required for alginate acetylation. J Bacteriol 175:5057–5065
40. Franklin MJ, Dennis EO (1996) Identification of *algI* and *algJ* in the *Pseudomonas aeruginosa* alginate biosynthetic gene cluster which are required for alginate O acetylation. J Bacteriol 178:2186–2195
41. DeLucca AJ II, Connick WJ, Fravel DR, Lewis JA, Bland JM (1990) The use of bacterial alginates to prepare biocontrol formulations. J Ind Microbiol 6:129–134