From Nano to Subnano Resolution: Reconstitution of Hydrofobin Structure from a SPM Scan of HFBI Layer

A. Paananen[1], V. Ruprecht[2*], M. Torkkeli[3], F. Kyriakopoulos[2], M. Müllner[2] and V. Pastushenko[2*]

[1] *VTT Technical Research Centre of Finland;* [2] *Johannes Kepler University of Linz;* [3] *University of Helsinki, Finland*

Abstract: Applying Scanning Probe Microscopy (SPM) methods to determine structural details of biomolecules (e.g. proteins) is one of the key trends in nanotechnology. Here we present the analysis of a Langmuir-Blodgett film of hydrofobin HFBI. The statistical analysis is based on an accurate procedure of data detrending to study the quasi-crystalline clusters of the HFBI layer. A statistical improvement method of signal to noise ratio was used. It relies on the variation of the crystallographic lattice superimposed on a quasi-periodical data set. The procedure produces optimal values of crystallographic periods and orientations of the symmetry axes. This information is used for selecting coherent data sets within a crystallographic cluster. Superposition of cells from the same cluster and calculation of an average periodical structure improves the signal to noise ratio, which leads from nano to subnano resolution of the protein structure. The number of superimposed cells N defines the resolution gain factor $N^{1/2}$. In the case of hydrofobin, the value of N lies in the range of 25-100. This allows for the determination of fine structural details of HFBI. Average structures obtained from different domains of the protein layer also show a high degree of reproducibility.

THE AIM

Surface layers of highly ordered proteinaceous structures of different bacteria including those living in non typical environments [1] gave rise to special research directions of S-layers [1-9]. If the structures are chemically dissolved, they occasionally show the ability to reassemble, sometimes even on a solid substrate surface [10]. This renders biotechnological applications of S-layer methodology possible [11], including their AFM studies [10-13]. In addition, Scanning Probe Microscopy with nano resolution has been used to resolve the structural details in Langmuir-Blogett films of highly ordered hydrofobin HFBI [14-16], a surface active protein produced by filamentous fungus *Trichoderma reesei* [17].

This paper is focused on the protein structure analysis of HFBI from SPM records. In particular, we present an optimization method for finding periods and crystallographic axes within a quasi periodic protein layer. The approach can be divided into two steps: 1) preprocessing of the raw SPM data, with the aim of removing some empirical artefacts, and 2) optimal detection of the periodical structure. The second step is essentially based on improving the signal to noise ratio by statistical methods based on averaging over several periodicity cells of the protein lattice. The superposition of a number of periodicity cells is straightforward if the optimal values of the periods are known. Therefore, step 2´s main emphasis is on the optimal detection of the lattice periods.

METHODS

Analysis of SPM data can be divided into three main steps:

1) Data preprocessing aiming to clean the data of different experimental artefacts.

2) Application of optimization methods for detecting periods and symmetry axis within selected crystallographic clusters.

3) Statistical analysis of the results for different crystallographic clusters.

I. Preprocessing

We illustrate the main preprocessing methods using SPM data of a HFBI layer [14]. Fig. (**1**) shows the raw data representing a HFBI layer on the surface of a flat substrate. Detilting (removing the slope) of the data can be made by the "surfit" method, c.f. Matlab File Exchange http://www.mathworks.com/matlabcentral/fileexchange/loadAuthor.do?objectId=1094148&objectType=author.

Fig. (**2**) shows the data, detilted by subtracting a least-square fitted plane. One can see some data clusters which are still not coplanar, because top and bottom of the figure have lower values than in the middle. Better flattening is achieved by a quadratic bipolynomial.

Fig. (**3**) shows the data after eliminating a parabolic cylinder obtained by "surfit" with degree 2 in y-direction and degree 0 in x-direction. This allows crystallographic clusters to be presented with much higher contrast. As the colorbar shows, the data within the HFBI film vary in a range of approximately 1 nm. The drift of the scanned rows is evident and is expressed in stripes around row 70 and row 180. Deep dark spots reflecting some substrate imperfections are scaled to about 2 nm. The spots depict the evident difficulty of eliminating the drift.

*Address correspondence to these author at the Johannes Kepler University of Linz, E-mail: verena_ruprecht@gmx.at
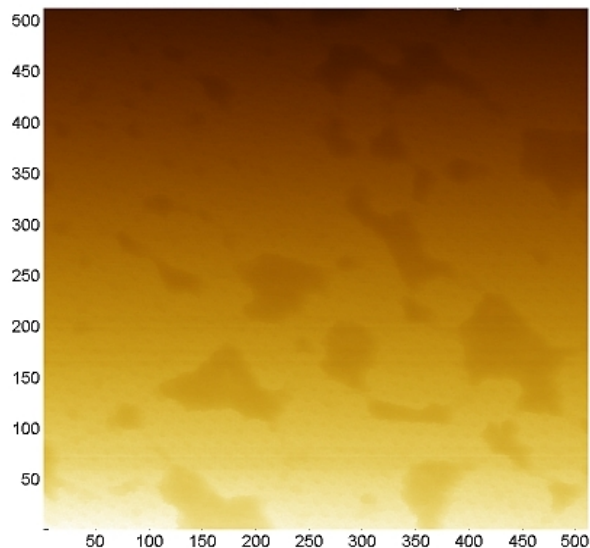
Fig. (1). Raw HFBI data (512 by 512 pixels, scan size 200 by 200 nm). Note a certain image tilt which is expressed by the signal decrease with growing row number.

Thus, if we eliminate the row drift by subtracting the average value, we get the result shown in Fig. (**4**). Narrow stripes occurring in figure 3 disappear, but shadow artefacts are now created due to large scale substrate imperfections represented as wide dark stripes in Fig. (**4**). To avoid this artefact, a special procedure of dedrifting was used. This is conveniently done using Matlab NaNotechnology (NaN means Not a Number) by setting values under a certain threshold to NaN. To discriminate between the signal and structural imperfections, it is informative to build a histogram of all data values shown in Fig. (**5A**). The signal (right peak) and the dark spots (left peak) are well separated, therefore the boundary between them is well defined and one can flatten the rows while ignoring the lattice imperfections. A new converged histogram Fig. (**5B**) can be obtained by repeating this procedure several times, which leads to well dedrifted data (Fig. **6**). Note that the range of data scattering
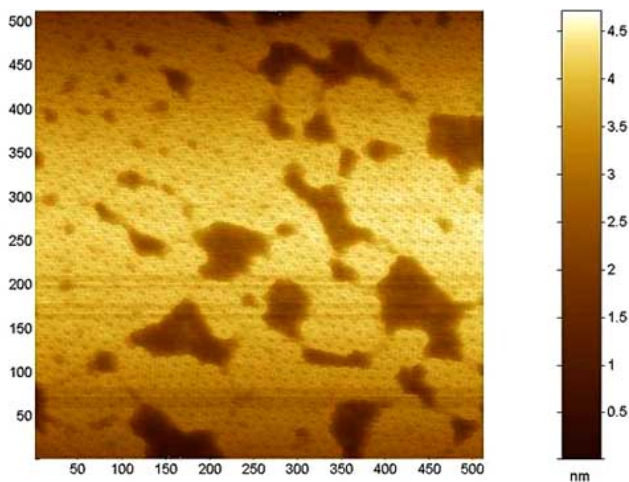
Fig. (2). Detilted raw data. Note the strong remaining deviation from flatness, shown by the height values scaled by the colorbar.
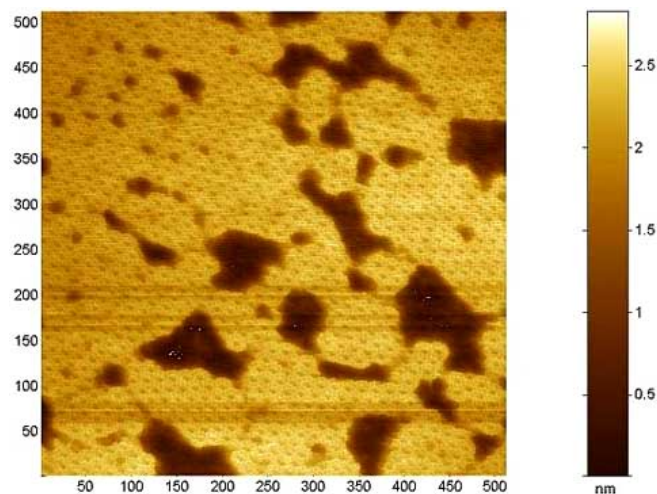
Fig. (3). Raw hydrofobin data flattened by a cylinder parabolic in y-direction.
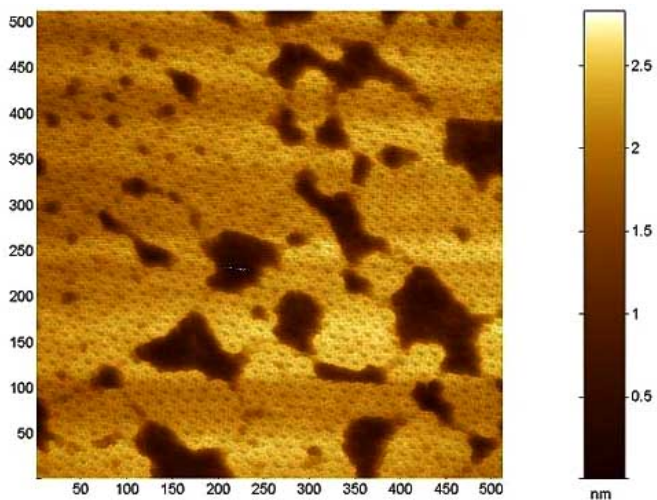

Fig. (4). Row dedrifting eliminates the jumps [··] between closely positioned data rows, but large-scale inhomogeneities produce undesirable shadow artefacts expressed as wide dark stripes.

has remarkably decreased, and the main peak corresponding to the signal values is much sharper than before.

After flattening and dedrifting the raw SPM data, different clusters (islands) of crystallographically consistent data sets were selected (Fig. **7**). These clusters are denoted according to their positions. Each data cluster was analyzed independently, both in terms of crystallographic periods and average structure.

II. Algorithms

The detection of periods, including the orientations of the symmetry axis in a given data cluster, is based on autocorrelation analysis. As a criterion for the reproducibility of structures belonging to the same cluster, the variance of the elementary structure collected from different cells was selected. Cells are arranged in a 3D array with the variances in all
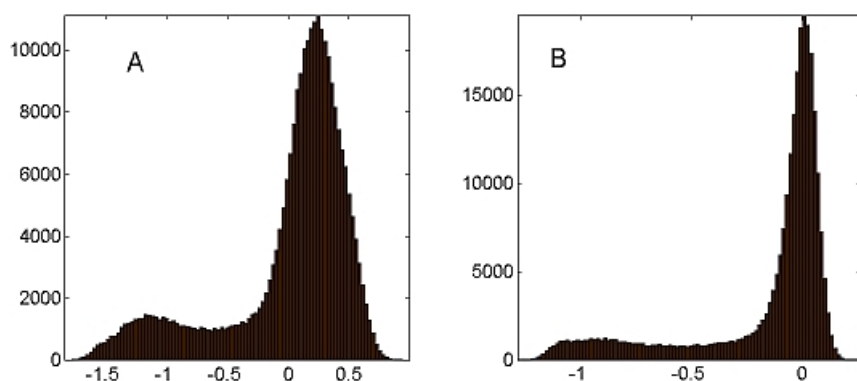
Fig. (5). Histograms of all data values after two-dimensional drift elimination: all points (A) and imperfections ignored (B). Improvement due to NaNo-flattening is expressed in a much sharper peak of data containing the signal in B.
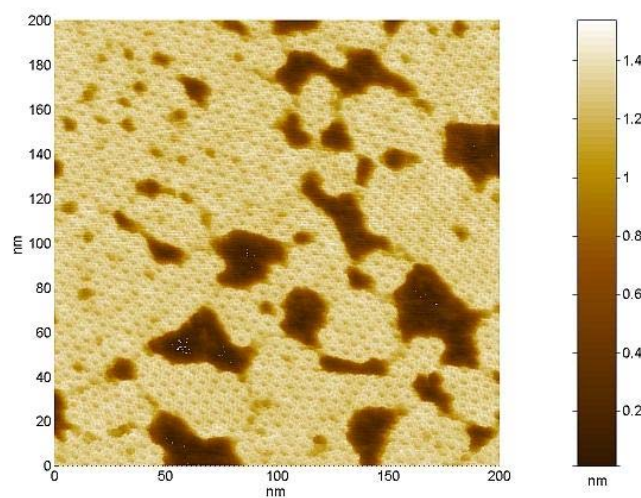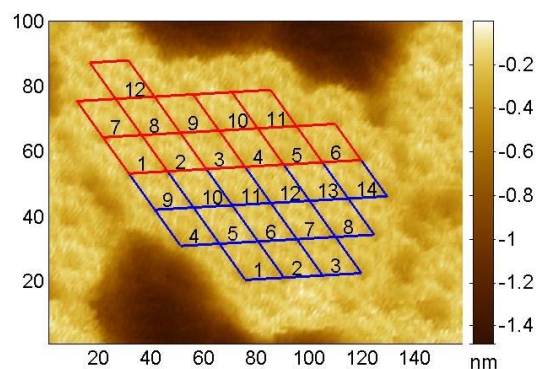

Fig. (6). Flat HFBI data produced by surfit and NaNo dedrifting of the scanned rows.
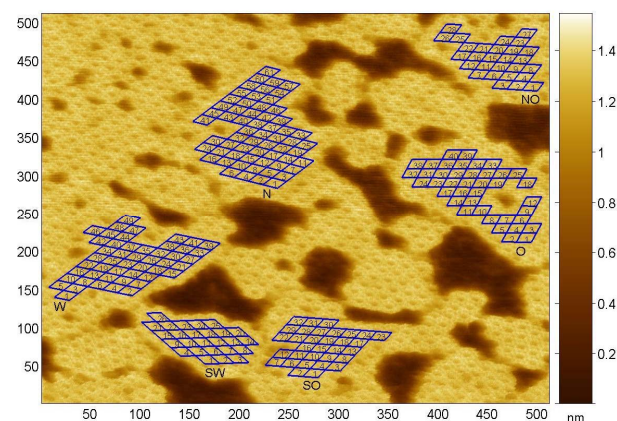
Fig. (7). Selection of crystallographically consistent clusters (SW, W, N, NO, O, SO).

pixels calculated in the 3rd dimension. The average of these variances is taken as a structure variance within a cluster. The calculated periods within a cluster minimize this variance.


Fig. (8). Higher (red) and lower (blue) parts of SW data. Separate analysis allows for a quantitative characterization of reproducibility.

After the periods have been found, the protein structure within a given cluster is calculated by averaging the local structures. The contributions of different cells are weighted by coefficients depending on their deviation from the average. This feature can also be used for automatic cluster extension. Cells with strongly deviating structures are completely excluded, while the cells with normal contents are accepted for the cluster extension. To compare reconstituted structures from different clusters, an additional numerical procedure was used which searches for the minimal discrepancy between two compared structures.

RESULTS

One of the data clusters (SW) is shown in Fig. (**8**). The lattice in this cluster was split into two parts, LOWer and HIGHer (in Fig. **8**) blue and red respectively). The optimal lattice periods were found independently in both parts, and then in the whole cluster. The results for lower and higher parts are shown in Fig. (**9**).

Fig. (**9**) shows the weighted HFBI patterns (A: lower, B: higher cells). The colorbar scales the heights from 0.02 nm to 0.22 nm. Both figures are very similar. The difference is the slight increase in noise in higher cells. Therefore, the apparent contrast seems to be slightly weaker in Fig. **9B**. However, all the structural details in both figures are very similar. Thus, we have a high degree of structure reproducibility within the SW cluster. In both cases, a quantitative
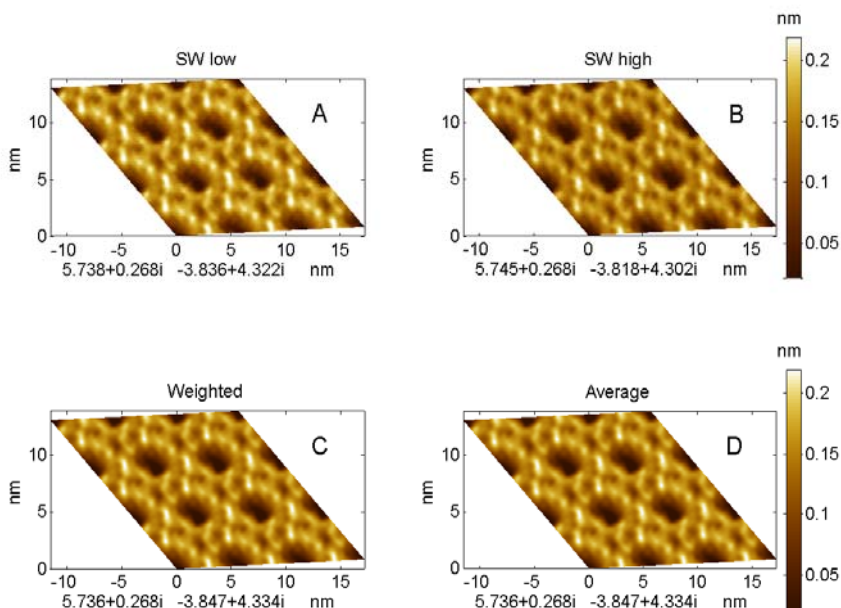
Fig. (9). Separate (A, B) and total (C, D) analysis in the SW data cluster. A, B: weighted patterns from lower and higher SW sublattices respectively. C: weighted average pattern. D: simply averaged pattern. The periods in nm are shown near the x-axis.
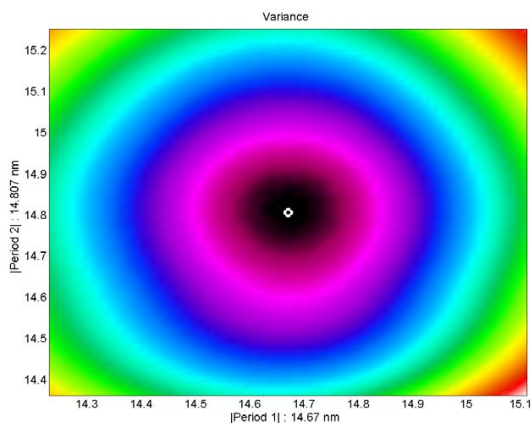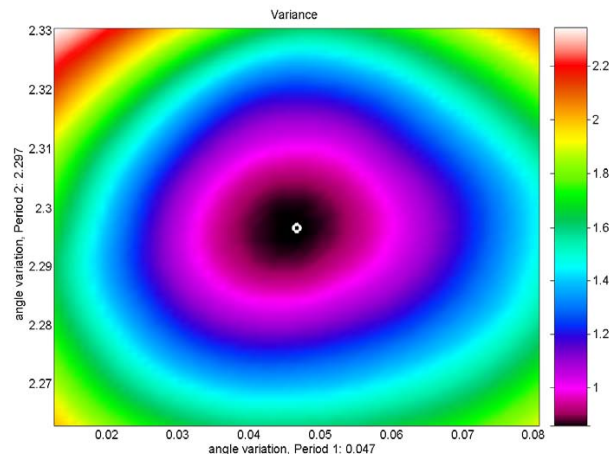

Fig. (10). Variance of SW cells as a function of the period values. Optimal absolute values of the periods are indicated near the axis.
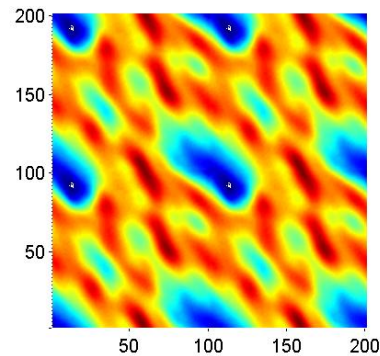

Fig. (11). SW variance dependent on period directions. Optimal angles (white circle) are indicated on the axis labels.

characterization of this reproducibility is given by period values shown below the x-axis. The difference between both versions is small.

Figs. (**9C, D**) summarize the results from the SW cluster. Fig. (**9C**) corresponds to weighted pattern, and **D** to mean pattern, constructed by simple averaging over all participating cells. The difference is weakly expressed indicating a high quality of both empirical data and optimal analysis.

Both Fig. (**10**) and Fig. (**11**) confirm the uniqueness of the minimum. After the periods were calculated in all data clusters, the average structures were constructed. The degree of periodicity can be checked by the presence or absence of jumps at the cell boundaries. This becomes clear if one repeats such structures in both directions. An example of doubling the structure is shown in Fig. (**12**).
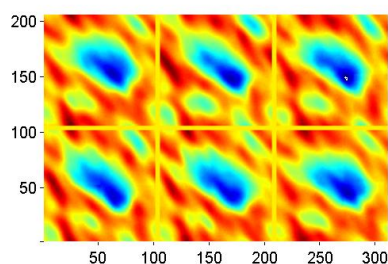

Fig. (12). Double average structure of SW cluster, consisting of four elementary cells.
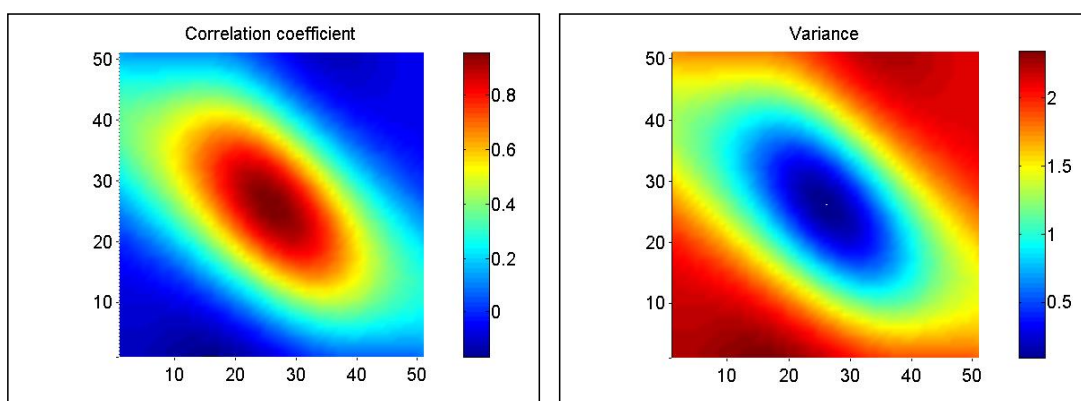
Fig. (13). Dependence of the correlation coefficient (left) and variance (right) between the SW and NO cluster on the position of the elementary structure.


Fig. (14). Optimal elementary average structures for all 6 islands standardized by comparison with the cluster NO.

For the purpose of cross-cluster comparison, an elementary average structure from one of the clusters is moved around the centre of a double structure from another island to find the optimal correspondence between the clusters (minimal cross variance or maximal correlation coefficient). Both criteria show the same extreme point.

As can be seen from Table **1**, the correlation coefficient, which serves as a quantitative measure for reproducibility between the islands, is very high (the minimum value is about 86%). The cluster NO primarily shows the highest correlation compared to other data islands. In Fig. (**14**) the standardized optimal structures (with respect to NO) are dis-

Table 1. Correlation Coefficient for Different Data Clusters

	SW	SO	W	N	O	NO
SW	1,0000	0,9502	0,9134	0,9214	0,9496	0,9631
SO	0,9502	1,0000	0,9719	0,9742	0,8774	0,9540
W	0,9134	0,9719	1,0000	0,9909	0,8597	0,9555
N	0,9214	0,9742	0,9909	1,0000	0,8656	0,9550
O	0,9496	0,8774	0,8599	0,8657	1,0000	0,9430
NO	0,9631	0,9540	0,9555	0,9550	0,9430	1,0000
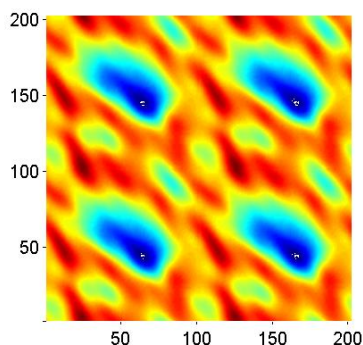
Fig. (15). Global doubled average hydrophobin structure from different cluster structures standardized with respect to NO.

played. These locally optimal average structures, standardized by comparison with the NO cluster, can be used for calculation of a global average structure (Fig. **15**).

SUMMARY

We suggest a procedure for deriving protein structure from periodical SPM data, based on selecting crystallographically consistent clusters from preliminarily flattened and dedrifted data by minimization of the structure variance. The method provides an optimal tool for extracting information from empirical data and is superior to conventionally used Fourier analysis, which is only applicable to rectangular data matrices.

REFERENCES

[1]   Sleytr, U.B.; Messner, P.; Pum, D.; Sára, M. *Mol. Microbiol.*, **1993**, *10*, 911.
[2]   Sleytr, U. B.; P. Messner. *Ann. Rev. Microbiol.*, **1983**, *37*, 311.
[3]   Sleytr, U.B.; Messner, P.; Pum, D.; Sára, M. *Mol. Microbiol.,* **1993**, *10*, 911.
[4]   Schuster, K.C.; Mayer, H.F.; Kieweg, R.; Hampel, W.A.; Sára, M. *Biotechnol. Bioeng.*., **2004**, *48* (1), 66.
[5]   Mengele, R.; Sumper, M. *J. Biol. Chem.,* **1992**, *267*, 8182.
[6]   Sleytr, U.B.; Sara, M.; Pum, D.; Scuster, B. *Prog. Surf. Sci.,* **2001**, *68*, 231.
[7]   Boot, H.J.; Kolen, C.P.H.M.; Pouwels, P.H. *J. Bacteriol.,* **1995**, *177*, 7222.
[8]   Egelseer, E.; Schocher, I.; Sara, M.; Sleytr, U.B. *J. Bacteriol.,* **1995**, *177*, 1444.

[9]   Sára, M.; Sleytr, U.B. *J. Bacteriol.,* **2000**, *182*(4), 859.
[10]   Beveridge, T.J. *Curr. Opin. Struct. Biol.,* **1994**, *4*, 204.
[11]   Sára, M.; Sleytr, U.B. *Micron*, **1996**, *27*,141.
[12]   Ebner, A.; Kienberger, F.; Huber, C.; Kamruzzahan, A.S.M.; Pastushenko, V.F.; Tang, J.; Kada, G.; Gruber, H.J.; Sleytr, U.B.; Sara, M.; Hinterdorfer, P. *ChemBioChem,* **2006**, *7*, 588.
[13]   Kienberger, F.; Pastushenko, V.F.; Kada, G.; Puntheeranurak, T.; Chtcheglova, L.; Riethmueller, C.; Rankl, C.; Ebner, A.; Hinterdorfer, P. *Ultramicroscopy*, **2006**, *106*, 822.
[14]   Paananen, A.; Vuorimaa, E.; Torkkeli, M.; Penttila, M.; Kauranen, M.; Ikkala, O.; Lemmetyinen, H.; Serimaa, R.; Linder*, M.B. Biochem.,* **2003**, *42*, 5253.
[15]   Hakanpää, J.; Paananen, A.; Askolin, S.; Nakari-Setäläa, T.; Parkkinen, T.; Penttilä, M.; Linder, M.B.; Rouvinen, J. *J. Biol. Chem.,* **2004**, *279* (1), 534.
[16]   Szilvay, G.R.; Paananen, A.; Laurikainen, K.; Vuorimaa, E.; Lemmetyinen, H.; Peltonen, J.; Linder*, M. B. Biochem.,* **2007**, *46*, 2345.
[17]   Linder, M.; Szilvay, G.; Nakari-Setälä, T.; Penttilä, M. *FEMS Microbiol. Rev.,* **2005**, *29*, 877.