Towards a Theory of Protein Adsorption: Predicting the Adsorption of Proteins on Surfaces Using a Piecewise Linear Model Validated Using the Biomolecular Adsorption Database

Dan V. Nicolau Jr[*], Dan V. Nicolau[†]

[*]Department of Mathematics and Statistics
University of Melbourne, Parkville 3010, Victoria

sarmisegetusa@yahoo.com

Industrial Research Institute Swinburne
Swinburne University of Technology, Hawthorn 3122 Victoria

dnicolau@swin.edu.au

Abstract

Predicting protein adsorption from solution to a surface is a perennial problem in biomedicine and related fields. Despite constant attention in the literature, it is not currently possible to predict quantitatively the amount of adsorbed protein given environment, protein and surface parameters. In previous work, we presented the Biomolecular Adsorption Database, an online collection of protein adsorption data collected from the literature, and more recently a program and set of algorithms for computing physico-chemical descriptors on protein surfaces. In this paper, we present a purely empirical approach to predicting protein adsorption using a linearly piecewise model with breakpoint. This model makes use of the previously developed surface property algorithms to describe the protein. We fitted and validated this model using the Biomolecular Adsorption Database. This model is capable of accounting for over 90% of the variance in the data, despite the fact that the adsorption data spans over three orders of magnitude. This represents a significant improvement over previous predictive modelling results.

Keywords: protein adsorption, database, modelling

1 Introduction

Protein adsorption is a very complex process and has for the most part resisted efforts to predict it quantitatively. Despite receiving constant attention, the general problem of predicting with any degree of accuracy the amount of adsorbed protein given protein, surface and solution parameters is far from being solved. This is due to several specific difficulties.

The most obvious of these is that the sheer complexity of the process is not captured by any current analytical model. Most approximate the protein by a rigid sphere of uniform charge. Neither of these simplifications is actually justified. In addition, no single model takes into account the large number of interactions and fluid-flow effects that affect the adsorption process.

In fact, at scales comparable to protein dimensions, the continuum hypothesis that underlies most theoretical models breaks down; at the same time, treating the problem accurately by computational chemistry (e.g. by molecular dynamics simulations) is at present not possible due to the large numbers of atoms involved.

Another problem is that the adsorption seems to be very sensitive to initial conditions in quite a few variables. Thus a small change in pH, temperature and ionic strength can make a large difference in the adsorbed amount. This also means one must be very careful when repeating an experiment, and hence that it is difficult to obtain reliable data.

On the other hand, it is important to know the protein concentration on the surface in many applications, such as biomaterials etc. for purely pragmatic reasons. To this end, we propose the empirical prediction of protein adsorption based on a piecewise linear model fitted to collected protein adsorption data. This differs significantly from our attempts to predict protein adsorption using a semi-empirical approach (Nicolau Jr. and Nicolau, 2002), which involves building physically sensible "candidate" models with undetermined parameters, and then determining these by fitting the model to the data.

2 Methods

2.1 Collection of Data

We have described the Biomolecular Adsorption Database (B.A.D.) in great detail in other work (Nicolau Jr. and Nicolau, 2002). In brief, this database is a

collection of protein adsorption data obtained from the literature. Although it is constantly growing, at the time of writing it contained over 400 entries. It is available online at www.bionanoeng.com/bad/.

Each entry in the B.A.D. lists the protein used, the surface to which it was adsorbed, the amount adsorbed, the bulk concentration of protein in solution, the surface tension or contact angle of the surface, the ionic strength of the solution, the buffer used, the temperature of the solution, its pH, the measurement principle used as well as a reference to the paper the data was obtained from.

From the entries in the database, we selected those for which an appropriate structural information file was available for the computation of the surface properties of the protein (see Section 2.2). In addition, we eliminated from this set the cases in which the adsorbed amount was unusually large, since they had a disproportionately large effect on the model parameters. It was judged that a compromise between finding an accurate predictive model for most adsorption scenarios without eliminating too many cases would result if adsorbed quantities above 4 $\mu g/cm^2$ were ignored. The final set used for the fitting of the piecewise linear model comprised 149 valid cases, representing data about 8 unrelated proteins (proteins whose functions are equivalent but which were obtained from different species are considered "related" here, e.g. bovine IgG and human IgG), namely α-chymotrypsin, α-lactalbumin, human growth hormone, fibronectin, β-lactoglobulin, lysozyme, ribonuclease, IgG and fibrinogen.

2.2 Calculation of Protein Surface Descriptors

The algorithms we have used to compute the properties of the molecular surface have been detailed elsewhere (Nicolau Jr. & Nicolau, 2002). We give here only a brief description of the methodology employed.

These algorithms are based on an extension of the well-known Connolly algorithm. The latter is described in Connolly (1983). The rationale behind this method of computation of the solvent-accessible surface of a molecule is that the only parts of the surface that are smaller than or around the same size as the solvent molecules actually interact with the latter: smaller "clefts" are not geometrically accessible.

Based on this idea, Connolly's algorithm rolls an imaginary sphere of a radius close to the effective radius of a solvent molecule (in the case of water, 1.4 Å) over the the 3-dimensional structure of the molecule under analysis, recording the contact points of this ball with the van der Waals spheres representing the constituent atoms of the latter. Together, these "pivot" points represent a good approximation to the solvent-accessible surface.

We extended this algorithm by assigning either to each atom or to each amino acid of a protein a real number representing some property of that atom or residue, for example charge or hydrophobicity. Then, while rolling the solvent sphere over the structure of the protein, we sum the values of this parameter of interest encountered at each pivot point on the surface. The result of this summation is an approximation to the integral of the parameter over the solvent-accessible surface. We implemented these algorithms in a freely available Windows™ program which we call the Protein Surface Properties Calculator (Nicolau Jr. and Nicolau, 2003).

A property such as charge or hydrophobicity cannot in most cases be meaningfully assigned to both individual atoms and individual residues. For example, most hydrophobicity scales use an amino-acid scheme while charge is considered to be a fundamental property of an atom rather than a set of atoms. However, the assignment of properties to entire residues, while possibly meaningful, presents some problems. Among these is the resulting assumption that the value of the parameter of interest is distributed homogenously across the residue's solvent accessible surface, which is not at all reasonable. Another drawback is that the best spatial resolution one can obtain is limited by the dimensions of a residue. These and other motivations led us recently (Nicolau Jr. and Nicolau, 2003) to propose a scheme for converting residue-based properties to atomic properties via a linearity approximation (i.e. the value of a parameter over a residue is equal to the sum of the unknown properties of the constituent atoms, possibly scaled by the exposed areas of these).

Although the assignment of physico-chemical parameters to individual atoms is not always physically meaningful (as is the case with hydrophobicity), it is more robust and can give numerically superior results. We showed (Nicolau Jr. and Nicolau, 2003) that mapping residue-based hydrophobicity scales to atomic ones leads to a more meaningful clustering of proteins and reveals atomic-resolution surface features that could not be determined by the amino-acid assignment alone. Consequently, in the present work we used this method to map the Kyte-Doolittle hydrophobicity scale to individual atoms based on a scheme involving 12 different atomic types (atoms were grouped by chemical similarity). It was this measure that was used to compute the surface of hydrophobicity and hydrophilicity.

To determine the charges on the atoms in each protein, we used semi-empirical methods on each amino acid terminated by two different identical residues at each end to determine the charges on the constituent atoms of the central amino acid. We make the assumption that the effect of more distant residue bonds and that due purely to different locations in a folded protein to be negligible. The resultant charges were used to compute surface charge properties for each protein in the data set.

The parameters we calculated for each of these proteins are listed in Table 1 with a brief description. Clearly, these are not all mathematically independent of each other, and some are even directly related or linear combinations of one another (e.g. the total area equals sum of the positively and negatively charged areas). Thus, we attempted to select an appropriate subset of parameters from this table to use for the final model fitting.

In addition to these, we also used the principal descriptors of the environment and adsorbing surface, available from the B.A.D.: protein bulk concentration in mg/ml, pH of the solution, ionic strength of the solution in mM and the surface tension of the adsorbing surface in dyne.cm. The isoelectric point of each protein was calculated using HyperChem. As the temperature used for the measurements in the final data set was almost always very close to room temperature, this variable was not included.

Symbol	Meaning
Area	Surface Area
PosArea	Positive Surface Area
PosCh	Total Positive Charge
NegArea	Negative Surface Area
NegCh	Total Negative Charge
TotCha	Total Charge at the Surface
HyPhiA	Hydrophilic Area
HyPhoA	Hydrophobic Area
TotHypho	Total Surface Hydrophobicity
TotHyPhi	Total Surface Hydrophilicity
SpPosA	Specific Positive Area in
PosChDen	Surface Positive Charge Density
SpPosChD	Specific Positive Charge Density
SpNegA	Specific Negative Area
NegChDen	Surface Negative Charge Density
SpNegChD	Specific Negative Charge Density
SpHphiA	Specific Hydrophilic Area
HphiDen	Hydrophilicity Surface Density
SpHphiD	Specific Hydrophilic Surface Density
SpHphoA	Specific Hydrophobic Area
HphoDen	Hydrophobicity Surface Density
SpHphoD	Specific Hydrophobic Surface Density

Table 1: Surface Parameter Calculated Using the Protein Surface Properties Calculator (areas are in Å^2 and "specific" means scaled by total surface area).

We obtained structure files for the proteins listed above from the PDB databank at www.pdb.org. We then used the Protein Surface Properties Calculator to compute the surface parameters listed in Table 2. The probe radius used was 100 Å. The reason for using this very large value instead of the customary 1.4 Å is that in the interaction of a protein with a surface (which is what controls adsorption) it is not the surface accessible to a water molecule which matters, but the surface accessible to an infinite plane representing the adsorbing surface. In the limit as $r \to \infty$, probing a protein structure with a sphere of radius r would clearly give the "plane-accessible" surface area of a molecule. The solvent-accessible and plane-accessible surfaces of a molecule are actually very different, as illustrated in Figure 1 below. We have found that surface parameters computed using large probe radii are more strongly correlated with adsorbed amounts than are those computed using smaller probe radii, and this difference represents around 30% of the correlation coefficient for most of those parameters.

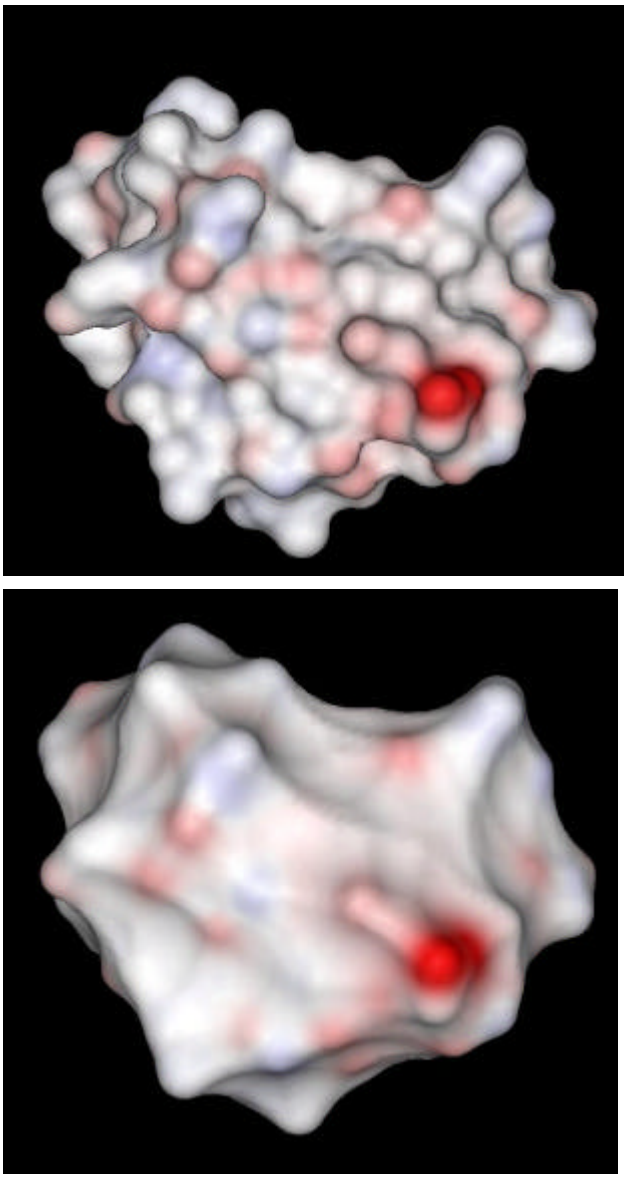

Figure 1. Water-accessible (above) and plane-accessible (below) surfaces of the same protein (crambin).

2.3 Piecewise Linear Model Fitting

We attempted to fit to the data a piecewise linear model with a breakpoint. This is in a sense the simplest non-linear model. Let the dependent variable be v_0 and the independent variables used in the model be $v_1, v_2, \ldots, v_n$.

Then according to a piecewise linear model with a single breakpoint b

$$\begin{cases} v_0 = \sum_{i=1}^{n} b_{1i} v_i, v_0 \le b \\ v_0 = \sum_{i=0}^{n} b_{2i} v_i, v_0 \ge b \end{cases} \quad (1)$$

where the b_{1i} are the parameters of the model for the region before the breakpoint i.e. $v_0 \le b$ and the b_{2i} are the corresponding parameters for $v_0 \ge b$. Additionally, we have the condition that the model must be continuous, i.e. at $v_0 = b$, both expressions take the same value.

In essence, this model works with the assumption that there are two different "regimes" for the dependence of v_0 on the other variables. Protein adsorption is known to be non-linear in time, bulk concentration and other variables. At low concentrations of protein in solution or under unfavourable conditions, protein adsorption is generally diffusion-limited. At higher concentrations and favourable conditions (e.g. high affinity of the protein for the adsorbing surface) the process is limited by the surface's finite capacity for adsorption. Hence there are strong reasons to suspect that the assumption of more than one regime of adsorption is well-founded, and thus that this model will be a much more appropriate one than a simple multiple linear regression.

We used the package Statistica™ with the data set defined as above and including the surface and environment parameters described in Section 2.2. This model was evaluated using a least-squares penalty function. The breakpoint is estimated manually, since we found that this gives better fits than allowing the algorithm to estimate it. It is the value at which the correlation between the observed and predicted values is highest. We used several different estimation algorithms: quasi-Newton, Hooke-Jeeves, Simplex, Rosenbrock and combination of Rosenbrock pattern search and quasi-Newton. Only the first and last of these actually converged to a satisfactory solution and both methods gave practically identical results. The maximum number of iterations was set to 1000 and the convergence criterion was set to 10^{-5} (the optimisation stops when the changes in the parameters from iteration to iteration are no more than the convergence criterion).

3 Results

As mentioned above, not all of the variables listed in Table 1 are independent of each other: in fact, most of them are not independent of the others. Thus we attempted to find a set of surface parameters which included surface area, charge descriptors and hydrophobicity descriptors but whose purely mathematical inter-dependencies were negligible. After some trials, we settled on a model comprising 10 independent parameters (including descriptors of the environment and surface). These are listed in Table 2, together with the final values of the parameters for both regimes of adsorption.

The optimisations required around 400 iterations to converge using both the quasi-Newton and Rosenberg quasi-Newton methods. In both cases the total residual over the 149 cases was 21.6, while the R2-value was 0.91. The optimal breakpoint was estimated to be around 1.75 μg/cm^2.

Coefficient of Parameter	Value before breakpoint	Value after breakpoint
Free term	9.020827	-0.79537
Bulk Concentration	0.169786	0.390651
Surface Tension	-0.04516	-0.01666
PH	-1.13612	-0.03771
Ionic Strength	-0.88226	-2.00472
Isoelectric Point	0.345245	0.376041
Surface Area	-3.5E-05	-2.5E-05
SpPosChD	125.4071	356.3409
SpNegChD	-312.176	-152.826
SpHphiD	-220.843	-142.063
SpHphoD	-707.82	-464.124

Table 2: Estimated coefficients for the parameters used in the model.

The plot of observed against predicted values is shown in Figure 2 below. Note the apparent "gap" near the area where the breakpoint was applied. Figure 3 shows the normal plot of the residuals. Note that these are close to normal (i.e. the plot is approximately linear).

The absence of any significant pattern about the "normal line" is an indication that the model is appropriate in the sense that no significant parameters have been omitted (as this would have caused the residuals to be non-normally distributed).

4 Discussion

Quite aside from the inherent physical complexity of the protein adsorption process itself, some of the parameters in the data set vary over several orders of magnitude. For example, the bulk concentration of protein in even the reduced data set used for fitting is as low as 0.001 mg/ml and as high as 10 mg/ml (4 orders of magnitude). The adsorbed amount and the ionic strength of the solution also show a similar degree of variability.

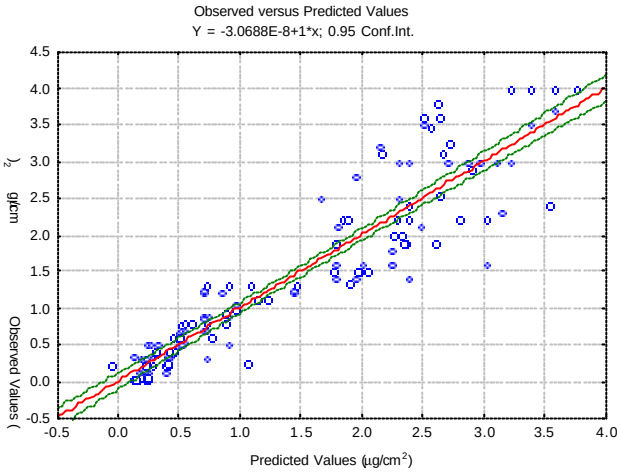

Figure 2. Observed vs. Predicted values of adsorbed amount in µg/cm^2. The dotted lines show the 95% prediction interval.

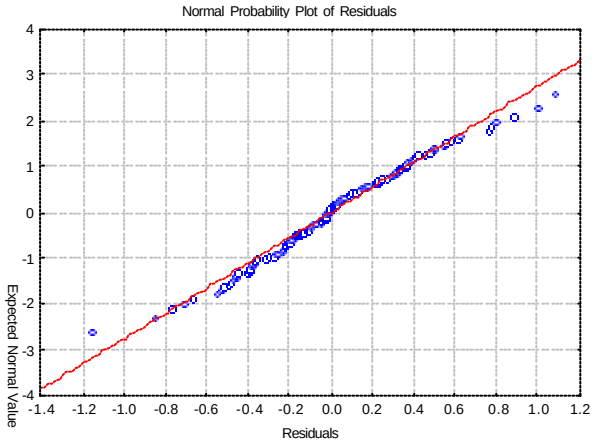

Figure 3. Normal probability plot of residuals. Note that the plot is almost linear.

Thus, it is not surprising that one linear regime is not enough to describe continuously the data in question. What is surprising to some extent is that only two linear regimes suffice to produce such a good fit. Given that the literature data is not always consistent and that realistically, the uncertainty in determining the concentration of protein at the surface is in most cases at least 10% if not more, it is quite surprising that any model can meet with such success, even a purely empirical one. Even if we relax the condition that the adsorbed amount be below 4 µg/cm^2, the piecewise linear model still explains no less than around 76% of the variation in the data while the optimal breakpoint is found in this case to be no higher than around 2.5 µg/cm^2.

The sufficiency of only two linear regions to obtain such a good fit for such a complex data set with such a high average uncertainty raises the interesting question of whether this optimal breakpoint has some physical significance or not. We can actually translate this into an approximate number of adsorbed layers using an approximate argument. Consider an "average" globular protein of nominal radius r and mass M. The projected area of such a protein, assuming it does not undergo any significant unfolding once adsorbed, is around πr^2. The densest packing of circles in the plane is the hexagonal "bee's honeycomb" lattice (Steinhaus, 1999), which has a packing density of around 0.907. Then, per unit area, one complete monolayer cannot have a mass of more than

$$\Gamma_{monolayer} \approx \frac{0.907M}{\pi r^2} \qquad (2)$$

If we substitute typical values into this approximate inequality, e.g. M = 50 kDa and r = 3 nm, we get values larger than 0.1 but below 1 µg/cm^2. In fact, proteins are known to unfold at the surface to some extent, packing is far from optimal, especially for non-globular proteins and the surface cannot always accommodate the highest density (the number of binding sites per unit area may limit its capacity to adsorb). Thus, the breakpoint of 1.75 µg/cm^2 would represent no less than a few monolayers at best. We are presently unable to explain this figure in physical terms.

We investigated whether varying the breakpoint would have a significant impact on the quality of the fit. The results are shown in Figure 4. Note that the difference between using breakpoints as low as 1.2 µg/cm^2 and as high as 2.4 µg/cm^2 compared to the optimal 1.75 µg/cm^2 is not great. This could be taken as an indication of some degree of "health" in the model, since it is clear that the breakpoint value is not due primarily to "artefacts" in the data. On the other hand, it is not clear whether this value represents something physically meaningful or not.

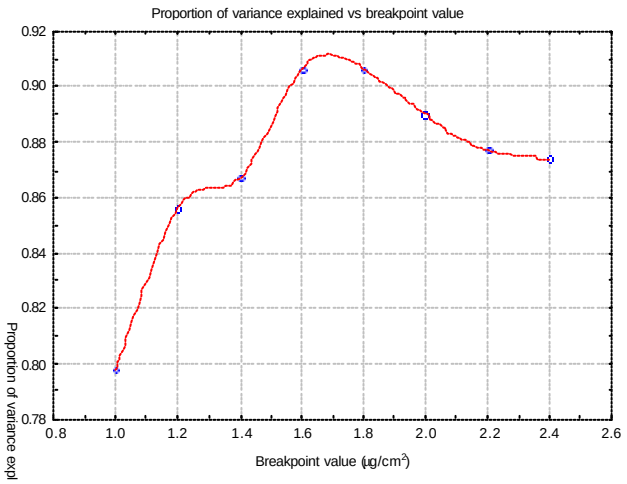

Figure 4. The variation of the fit of the model with a manually set breakpoint value (note the maximum around 1.75 µg/cm^2)

The use of an empirical model such as this does not in general have any theoretical implications for a phenomenon. Since we have used a simple non-linear model, it may be that some feature of the general data on protein adsorption found in the literature has been

captured but overall it is not clear that any new theoretical insights are provided by the values of the parameters as such (although this is possible). For example, it is intuitively obvious why the coefficients of the bulk concentration terms are positive: the higher the bulk concentration, the higher the protein adsorption, all things being equal. It is not so clear, however, why the coefficients of the surface area terms should be negative, for example.

It is also difficult to place this model in the context of other predictive efforts because to the best of the authors' knowledge, no such effort has yet been undertaken. Indeed, before the Biomolecular Adsorption Database, no collection of data from the literature of protein adsorption existed. Certainly, a great deal of effort has been directed at developing phenomenological models of biomolecular adsorption but no empirical analysis of the available data has even been attempted, much less been successful. Additionally, work on modelling the adsorption of proteins to surfaces is often not verified using any empirical data at all; when this is done, typically the data set consists only one or two proteins adsorbed in no more than a few conditions. In a recent paper (Nicolau Jr. and Nicolau, 2003) we have proposed a non-linear model for protein adsorption — this model had numerically inferior results (it was able to explain only around 75-80% of the variance in the data) but had the distinct advantage of being, in some sense, physically meaningful (although the coefficients were determined empirically). Continued work on this topic will certainly shed some light on the problem of protein adsorption and will also provide a context

On the other hand, the ability to predict protein adsorption to surfaces from solution over such a wide range of conditions, proteins and surfaces is very useful for pragmatic reasons and finds applications in biomedicine and other areas.

5 Conclusion

We have collected literature data on protein adsorption to solid surfaces from solution and used a piecewise linear model with breakpoint to describe it. We find that this type of model produces a very good fit and is capable in a least-squares sense of explaining over 90% of the variation in the data.

6 References

Nicolau, DV Jr. and Nicolau, DV (2003): A New Program to Compute the Surface Properties of Biomolecules, Proceedings of the 1st Australasian Conference on Bioinformatics

Nicolau DV Jr. and Nicolau DV (2002): A Database Comprising Biomolecular Descriptors Relevant to Protein Adsorption on Microarray Surfaces, *SPIE Proceedings*, Vol. 3, **18:** 109-116

Nicolau, DV Jr. & Nicolau DV. (2002): A Model of Protein Adsorption to Solid Surfaces from Solution, SPIE Proceedings, Vol 3, **18**: 1-9

Nicolau, DV Jr., Biomolecular Adsorption Database, www.bionanoeng.com/bad/.

J.L.Brash and T.A.Horbett Eds. (1987): Proteins at Interfaces Physicochemical and Biochemical Studies. ACS Symposium Series 343, ACS 1987.

T.A.Horbett and J.L.Brash Eds. (1995): Proteins at Interfaces II" ACS Symposium Series 602, ACS 1995.

Andrade, J.D. (1985) in Surface and Interfacial Aspects of Bio-medical Polymers (Andrade, J. D., Ed.), Vol. **2**: 1–80, Plenum Press, New York

Luo, Q. and Andrade, J.D. (1998) *J. Colloid & Interface Science* **200**: 104–113.

M. R. Oberholzer, A. M. Lenhoff (1999): *Langmuir* **15**: 3905-3914.

Steinhaus, H. (1999): Mathematical Snapshots, 3rd ed. New York: Dover, **3**: 202