Connecting Nanoscale Images of Proteins with Their Genetic Sequences

Brian A. Todd, Jayan Rammohan, and Steven J. Eppell
Department of Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio 44106-7207

ABSTRACT We present a technique for reconstructing biomolecular structures from scanning force microscope data. The technique works by iteratively refining model molecules by comparison of simulated and experimental images. It can remove instrument artifacts to yield accurate dimensional measurements from tip-broadened data. The result of the reconstruction is a model that can be chosen to include the physically significant parameters for the system at hand. We demonstrate this by reconstructing scanning force microscope images of the cartilage proteoglycan aggrecan. By explicitly including the protein backbone in the model, we are able to associate measured three-dimensional structures with sites in the protein primary structure. The distribution of aggrecan core protein lengths that we measure suggests that 48% of aggrecan molecules found in vivo have been partially catabolized at either the E(1480)-(1481)G or E(1667)-(1668)G aggrecanase cleavage site.

INTRODUCTION

One of the fundamental problems in biophysics is measuring the three-dimensional structures of important molecules. X-ray diffraction (XRD) (Abola et al., 2000; Lamzin and Perrakis, 2000) and nuclear magnetic resonance spectroscopy (NMR) (Montelione et al., 2000) are the traditional and highest resolution techniques available for this task (~1–5 Å). Their applicability, however, is limited to molecules that crystallize, in the case of XRD, and to those that produce relatively clean spectra, in the case of NMR. Scanning force microscopy (SFM) (Binnig et al., 1986) is a newer versatile technique for measuring three-dimensional biomolecular structures (Czajkowsky and Shao, 1998), albeit at substantially lower resolution (~1–10 nm).

SFM images of biomolecules almost never give atomic resolution, however, they provide a wealth of information on a >1 nm scale. This is illustrated using SFM images of the four major structural molecules from cartilage (Fig. 1). For instance, flexibility of the collagen backbone, reflected in the tortuous conformation of the type II collagen molecule (Fig. 1 *a*) is not generally observed by XRD or NMR. SFM also reveals other submolecular features, such as the distinct extended and globular domains seen in images of types XI (Fig. 1 *b*) and IX (Fig. 1 *c*) collagen. Likewise, attached polysaccharides on the proteoglycan aggrecan (Fig. 1 *d*) are evident in its distinctly larger width compared to the collagens (Fig. 1, *a–c*).

A significant difficulty arises when relating the morphological structures measured by SFM with other biophysical information that is chemically or biologically specific. As an example, we show in Fig. 2 a schematic representation of the amino acid sequence for bovine aggrecan, derived from cDNA analysis (Hering et al., 1997). Made explicit in this figure are regions where globular domains (*black ovals*) and extended domains (*white boxes*) appear, sites where glycosaminoglycans attach (*black lines extending from the core*), and five sites along the protein where aggrecan is cleaved by a class of enzymes called aggrecanase (*arrows*) (Tortorella et al., 1999). One of the main problems addressed in this article is how to relate this primary structure information, obtained by cDNA analysis (Fig. 2), with our SFM images of aggrecan (Fig. 1 *d*). For instance, we might ask, ''What is the three-dimensional structure associated with each of the aggrecanase cleavage sites?'' An SFM image of aggrecan like that in Fig. 1 *d* contains information regarding the three-dimensional structure of the molecule. However, there is no straightforward way of knowing, using only the SFM image, where the cleavage sites are. We cannot resolve the detailed atomic structure or make out the specific epitopes.

An additional complicating factor is that raw SFM images of biomolecules do not provide good quantitative measures of molecular structure (Eppell et al., 1993; Wilson et al., 1996). This is especially true for lateral dimensions, which are generally much larger than the actual dimensions of the biomolecule because of ''tip broadening''. This is an artifact whereby the size and shape of the SFM tip contribute to the measured width of a molecule. For example, using a fairly sharp SFM probe (calibrated apex radius ~6 nm), we obtained a width for the type II collagen in Fig. 1 *a* of 8 nm. This compared to a diameter of 1.3 nm measured by XRD of model peptides (Bella et al., 1994). Various treatments have been proposed to deal with this inaccuracy (Eppell et al., 1993; Keller, 1991; Markiewicz and Goh, 1994; Wilson et al., 1996). However, all of these methods suffer from the fact that there are regions of the molecular surface that are never probed by the SFM tip. As a result, these methods place an upper limit on the width and length of the molecule but often fall far short of yielding the true size of the molecule.

In this article, we describe a new method that removes inaccuracies associated with the finite size of the SFM probe and has the potential to yield the true size of an imaged molecule. The technique is in many ways analogous to the reconstruction techniques applied to XRD (Lamzin and Perrakis, 2000), NMR (Montelione et al., 2000), and

Submitted July 18, 2002, and accepted for publication January 28, 2003.
Address reprint requests to Steven J. Eppell, 10900 Euclid Ave., Cleveland, OH 44106. Tel.: 216-368-4067; Fax: 216-368-4969; E-mail: sje@cwru.edu.

0006-3495/03/06/3982/10 $2.00

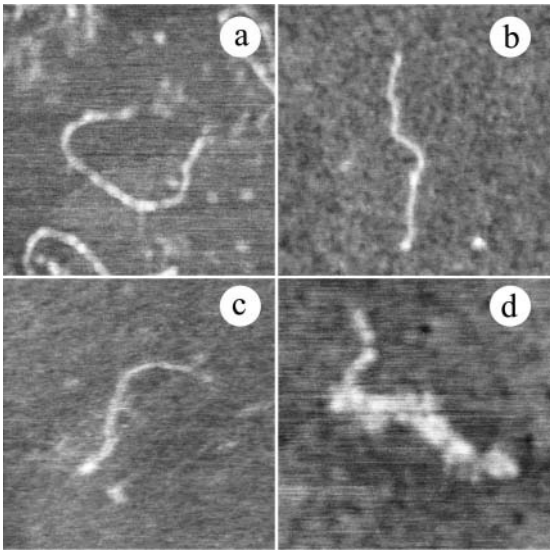

FIGURE 1 Typical scanning force microscope images of the four major structural proteins in cartilage. All images are 300 x 300 nm^2 with a 1 nm linear grayscale. (*a*) Type II collagen has a uniform flexible structure. (*b*) Type XI collagen contains both collagenous and a terminal globular domain seen near the bottom of the image. (*c*) Type IX collagen is a fibril associated collagen with interrupted triple helices. The interruptions appear as an elliptical amino-terminal globular domain (toward *left side* of the image) and also as ''kinks'' in the molecule. (*d*) The major cartilaginous proteoglycan aggrecan. The extensive glycosylation (~90% by mass) can be seen in the large and nonuniform width of this molecule.

transmission electron microscopy (Ernst and Ruhle, 1997) in that experimental data is used as a constraint to determine a best-fit structure for a model of the molecule. For SFM, this technique has advantages over previous reconstruction techniques in that it allows a priori information based on other biophysical measurements to be incorporated into the reconstruction of molecular structure. This increases the connectivity between SFM and other biophysical techniques. We demonstrate this by mapping tertiary structural features observed in SFM images of aggrecan onto primary structural locations of the protein known from cDNA (Hering et al., 1997). This links the three-dimensional structure of the protein to biochemical function via its genetic sequence.

THE RECONSTRUCTION METHOD

The method we have developed is describable in two parts: the reconstruction problem and models of nanoscale molecular structure. The former involves development of a mathematical process for reconstructing the underlying structures from an experimental SFM image. The latter involves developing models to describe biomolecular structures at the nanoscale.

Reconstruction framework

Any measurement process can be described by the equation,

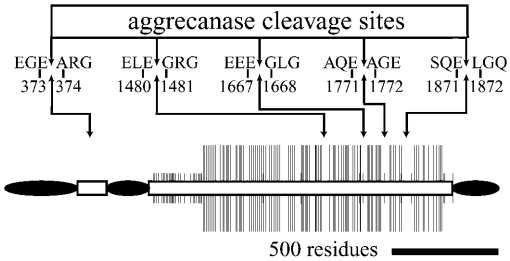

FIGURE 2 A schematic diagram for the cDNA sequence of aggrecan (Hering et al., 1997). In the model, globular domains are indicated by black ovals and extended domains as boxes with black border and white center. Lengths in the model correspond to number of amino acids (not distances in space). Aggrecan is known to be variably substituted with glycosaminoglycans along its protein core. These substitution sites are inferred from the location of attachment sites in the primary structure of the protein (Block et al., 1992; Hascall and Sajdera, 1970; Heinegard and Axelsson, 1977; Hering et al., 1997). Sites where keratin sulfate attaches (serine or threonine) to the backbone are indicated by short black lines running perpendicular to the backbone, and sites where chondroitin sulfate chains attach (serine-glycine tandem) are indicated by longer black lines. The five known aggrecanase cleavage sites at E(373)-(374)A, E(1480)-(1481)G, E(1667)-(1668)G, E(1771)-(1772)G, and E(1871)-(1872)L are indicated with arrows (Tortorella et al., 2000).

$$y = \Phi(x), \tag{1}$$

where an instrument $\Phi(\,)$ operates on a physical quantity x to yield a set of observables y. In the case of SFM, x is the three-dimensional structure of the surface under study (say a biomolecule adsorbed on mica) and $\Phi(x)$ describes how the SFM tip blurs the surface to produce an image y. If the instrument $\Phi(x)$ does not distort the surface structure too much, the image is very nearly the same as the surface ($y \approx x$), and we might be willing to accept y as a reasonable approximation of x. This is what is implicitly done by interpreting SFM images directly. However, for feature sizes below the characteristic dimensions of the tip, this assumption is generally not valid and we would do better to take into account the relationship between the image and the underlying surface structure that produced it. In particular, if we now think of $\Phi(x)$ as representing a mathematical model for our instrument (Villarrubia, 1997), we seek the inverse of $\Phi(x)$ that will recover the underlying physical quantities of interest from the image,

$$x = \Phi^{-1}(y). \tag{2}$$

The primary difficulty in reconstructing SFM images is that the inverse of the instrument operator (morphological dilation) does not exist (Villarrubia, 1997); Eq. 2 cannot be solved uniquely.

The approach we take to overcome this problem is to use Eq. 1 to check the consistency of some proposed molecular structure with experimental data. In other words, we proffer some molecular structure x^0 and then simulate the imaging process using Eq. 1 to obtain a simulated SFM image y^0. We then compare the simulated image y^0 to the experimental

image y to evaluate the quality of the proposed molecular structure. If we then iteratively adjust our estimates x^k using nonlinear regression, we can converge to a best-fit x^*. This best-fit represents our estimate of the structure of the molecule based on experimental measurements. It is ''optimal'' in the sense that the quantity,

$$\chi^2 = \sum_i \sum_j (y^*_{ij} - y_{ij})^2, \quad (3)$$

the summed squares of the pixel-by-pixel differences between the simulated and experimental image, is minimized.

An example that illustrates the important quantities in the nonlinear regression loop is shown in Fig. 3. The top image shows a proposed structure for a biomolecule adsorbed on mica and the calibrated (Todd and Eppell, 2001) SFM tip (*gray*) that was used to simulate (Villarrubia, 1997) the SFM image. The simulated image can be compared with the experimental image of the molecule by looking at their pixel-by-pixel difference. That the difference image is essentially indistinguishable from the noise confirms quantitatively what can be seen by eye; the simulated and experimental images are very similar. Note, however, that the reconstructed surface representing the actual structure of the biomolecule (*model surface*) has significantly smaller dimensions than the experimental image. This is a consequence of the tip broadening effect and highlights the importance of image reconstruction.

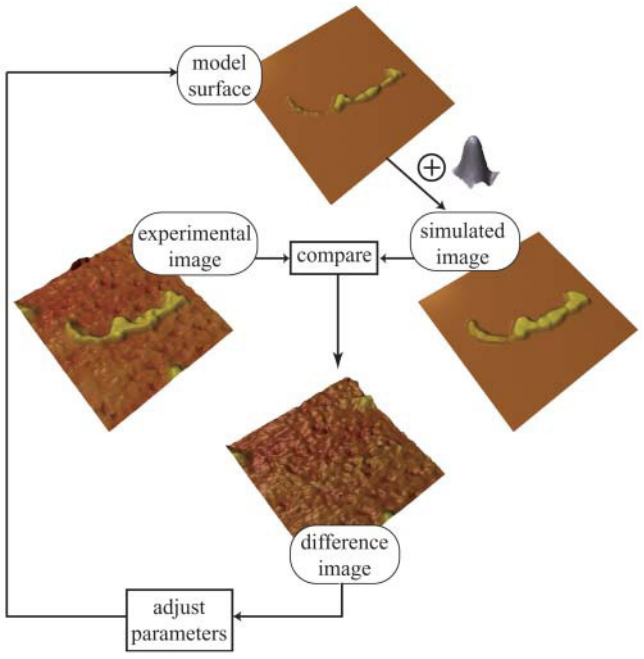

FIGURE 3 A graphic representation of the nonlinear regression loop used to extract molecular models from SFM images. A parameterized proposed model surface is dilated using the experimentally determined SFM probe shape (*gray*) to produce a simulated image. The simulated image is then compared to the experimental image to produce a difference image. The difference image is summed pixel-by-pixel to quantify the goodness-of-fit and direct adjustment of the model's parameters.

Nanoscale biomolecular models

To use a nonlinear regression approach for image reconstruction, it is necessary to describe the surface structure in terms of a model with parameters that can be adjusted to fit the data. The resolution in SFM is not sufficient to use atomistic models (as is done with XRD and NMR), so it is necessary to develop phenomenological models with nanoscale features. For fibrillar proteins, a model based on a geometric primitive used in medical imaging called a generalized cylinder (Harris and Stocker, 1998) yields a facile representation of protein structure. An example of this model is shown in Fig. 4. The model uses parametric cubic-splined curves (*white line*) that are defined by a set of control points or ''knots'' (*red spots*) to describe the orientation of the major axis of aggrecan in the plane of the substrate. Geometric aspects of the molecule (width, curvature, etc.) referenced to this contour line can be related to primary sequence via Fig. 2. Heights and widths perpendicular to the major axis are controlled by a set of elliptical cross sections (*yellow hoops*) with independent major and minor axes that are interpolated along the backbone to sweep out full three-dimensional structures (seen as the translucent shell of the model). We use these to model the nonuniform glycosylation pattern along the protein core of aggrecan (see width of the molecule in Fig. 1 *d*). Conceptually, one might think of the model as similar to a snake that has swallowed eggs. By adjusting the parameters, the generalized cylinder is capable of modeling all of the proteins in Fig. 1. We stress that the model obtained from this process is not simply a collection of image pixels, but a well-behaved function. Because of this, it is amenable to differentiation, integration, and a host of other operations useful for quantifying protein structure.

TESTING THE RECONSTRUCTION METHOD BY SIMULATING SFM IMAGES OF AGGRECAN

We wished to test the consistency and accuracy of results obtained using our new method. Since the general idea of using mathematical morphology to model SFM imaging is well accepted, we did not think it sufficient to test our method using a simple calibration standard. Instead, we spe-

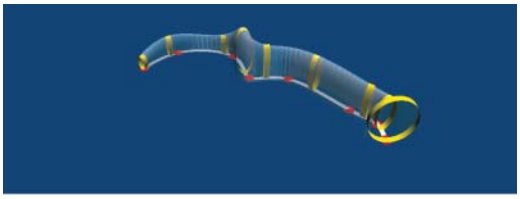

FIGURE 4 A generalized cylinder is used to model the molecular structure of aggrecan. The protein backbone (*white line*) is a spline parameterized by 10 ''knots'' (*red dots*). The width and height of the molecule is controlled by a set of ellipses (*yellow hoops*). The hoops are interpolated along the backbone (shown as the translucent shell) to give the complete three-dimensional surface of the molecule.

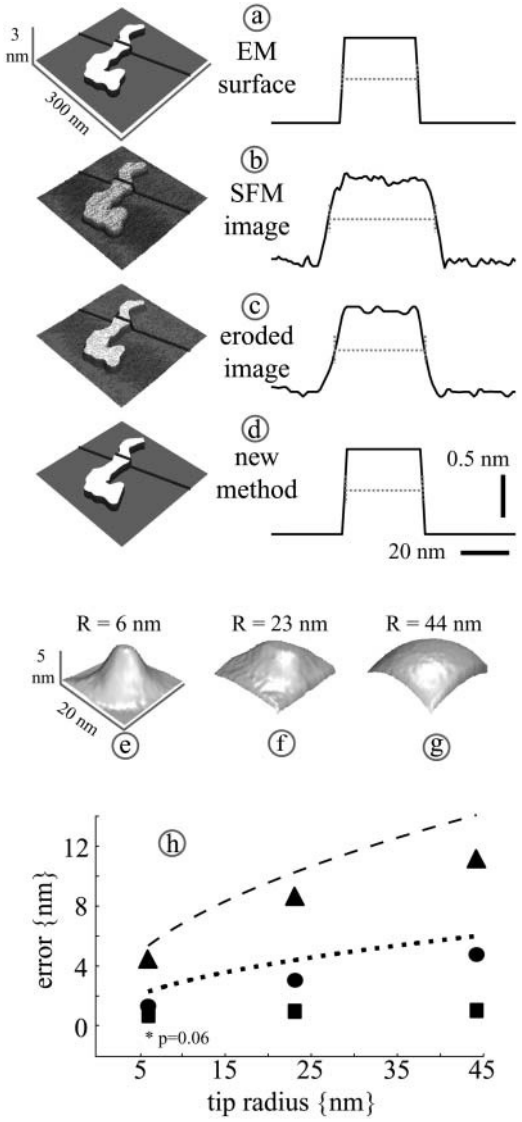

FIGURE 5 Comparing the new reconstruction method with previous techniques. (*a*) One example of the 23 molecules generated from electron micrographs of aggrecan molecules absorbed on mica (Buckwalter et al., 1985). A cross section through the molecule (*solid black line*) is used as a standard of comparison for the accuracy of each technique. (*b*) SFM image simulated using a previously described technique (Villarrubia, 1997) and an SFM tip with radius 44 nm (*g*). Lateral dimensions in the SFM image are broadened compared to the true dimensions from the EM image by 12 nm. (*c*) Reconstruction by erosion yields an improvement over the SFM image but the lateral dimensions are still broadened by 5 nm compared to the EM surface. (*d*) The image obtained using the new method. The lateral dimension of the molecule is the same as the TEM surface within the 1 nm pixel size of the original SFM tip image (mathematically computed error of 0.3 nm). (*e–g*) Experimentally calibrated SFM probe tips with approximate radii of 6, 23, and 44 nm, respectively. (*h*) Plots of the average errors in measured lateral dimensions as a function of the size of the SFM probe used to collect the image. Three processes are explored to extract the lateral measurements: direct measurement from the raw SFM image (▲), measurement from the eroded image (●), and measurement from the best-fit model obtained using the new method (■). Each data point in the figure represents the mean errors for five cross sections taken at distinct locations along a given molecule (23 different molecules were used giving a total of $N = 115$). The error in the SFM image increases with probe radius roughly according to $2\sqrt{R}$ (*dashed line*). The error in the eroded image increases with probe radius roughly according to $0.8\sqrt{R}$ (*dotted line*). The errors in the new method were smaller than the pixel size of 1 nm regardless of the probe size. The reduction in error obtained using the new method was statistically significant with $p < 0.0001$ in all cases, with the exception of the smallest probe radius (denoted *), where $p = 0.06$.

cifically tested the ability of the method to recover ''known'' dimensions of an object with complex shape. We used structures of aggrecan measured previously by transmission electron microscopy (TEM) as our standard. Rotary shadowed replicas of aggrecan molecules from published EM images (Buckwalter et al., 1985) were used to construct topographs representing the surface of aggrecan absorbed on mica. Simulated SFM images were created using three different experimentally determined probe geometries as previously described (Villarrubia, 1997; Wilson et al., 1996). These simulated images were used as inputs to the reconstruction method. Comparison of the output of the method with the original TEM data allowed a direct evaluation of the consistency and accuracy of the new method.

Lateral dimensions of aggrecan monomers were derived from TEM images by digitizing 23 nonaggregated monomers from Fig. 1 of Buckwalter et al. (1985) using a flatbed scanner. The image was scanned in at 1200 dpi (~0.66 nm/pixel) using an HP Scanjet IIcx (Hewlett-Packard, Palo Alto, CA). Regions of interest containing individual molecules were extracted from this image, exported to a separate file, and scaled to give a final pixel resolution of 1 nm/pixel (commensurate with the pixel size of the SFM tip images). The two-dimensional TEM images were converted into three-dimensional objects by attributing a nominal height of 1 nm to each molecule. An example is shown in Fig. 5 *a*. Gold standards for width measurements were obtained by taking the full width at half-max cross section (*gray dotted line* in Fig. 5 *a*). This is the same measurement one would obtain by simply laying a ruler across the original TEM image and measuring the width of the molecule. The benefit of taking the full width at half-max measurement is that we were able to directly compare exactly the same line in the image using the same computer code to make all measurements. A simulated SFM image of the molecule was generated (Villarrubia, 1997) by dilating the surface with an experimentally calibrated (Todd and Eppell, 2001) probe shape (Fig. 5 *g*) and adding a noise image obtained from an SFM scan of a clean mica surface (Fig. 5 *b*). The cross section of the molecule is increased in the SFM image compared to the TEM surface by ~12 nm. This represents the error expected when measuring the width directly from an SFM image of aggrecan using a tip with effective radius 44 nm. We applied the erosion reconstruction technique (Villarrubia, 1997; Wilson et al., 1995) to obtain the topograph in Fig. 5 *c*. Lateral distortions are substantially reduced; however, a residual error in the width of ~5 nm remains. This represents the error expected when measuring

the width of an aggrecan molecule from an eroded SFM image (the current state-of-the-art method) obtained with a 44-nm radius tip. Using the new method, we reconstruct the molecule shown in Fig. 5 *d*. The structure closely matches the original EM surface. The cross section measurement obtained by the new method matches the TEM gold standard within the limits set by the digitizing errors of the original SFM tip image.

To make a statistically significant comparison of the methods, the average errors for width measurements obtained for five different cross-sections through 23 different molecules ($N = 115$) were calculated. This process was repeated using experimentally calibrated (Todd and Eppell, 2001) geometries for three different probes (Figs. 5, *e–g*) of effective radius 6, 23, and 44 nm (radii obtained by fitting a sphere to each tip). The mean errors for the SFM image (▲), the eroded image (●), and the new method (■) are shown as a function of effective radii. As expected, the average errors in the raw SFM images were the largest. The increase in error as a function of tip radius is similar to a $\sqrt{4R - 1} \approx 2\sqrt{R}$ dependence (*dashed line*) expected based on a simple geometric analysis of a circle of radius R dilating a box 1 nm high. The erosion reconstruction improves on this, reducing the error by 60%, on average. The residual errors are again very close to those expected based on a simple geometric analysis of $2\sqrt{2R - 1} - \sqrt{4R - 1} \approx 0.8\sqrt{R}$ (*dotted line*). The new method had an average error of 0.3 nm, with no significant change with tip radius ($p > 0.01$ by Student's *t*-test). In all cases but one, the improvement in mean error using the new method was statistically significant, with $p < 0.00001$. In the one case of the smallest tip of radius 6 nm, the significance in the improvement of the new method over erosion was marginally different with $p = 0.06$.

EXPERIMENTAL METHODS FOR SFM IMAGING OF AGGRECAN

Materials

Aggrecan was obtained from 1–2-year-old bovine metacarpal-phalangeal articular cartilage purified under associative/dissociative conditions (A1A1D1 fraction) (Rosenberg et al., 1991). This purification procedure resulted in a population of aggrecan molecules that all contain the amino-terminal end and some fraction of the remaining length, depending on the extent of degradation.

SFM imaging

All images were obtained using a Nanoscope III Multimode SFM (Digital Instruments, Santa Barbara, CA) in air tapping mode. The geometry of the Pointprobe (Nanosensor, Norderfriedrichskoog, Germany) SFM probe used in the experiment was characterized by a modified blind reconstruction method (Todd and Eppell, 2001) from images of a Nioprobe tip characterizer (General Microdevices, Edmonton, Alberta, Canada). Images of the tip characterizer were obtained using 128×128 nm^2 fields-of-view with 256×256 pixels. Before reconstructing the tip, the pixel density was reduced to 128×128 by mean averaging 2×2 clusters of pixels to obtain a single pixel every 1 nm. This reduced the overall noise level in the image by a factor of $\sim$2, but more importantly, reduced the scan-direction dependent anisotropies in the noise that corrupt the blind reconstruction process (Todd and Eppell, 2001). Protein samples were prepared for imaging by depositing 5 μl of protein in 10 mM ammonium acetate buffer at a concentration of 1 μg/ml onto freshly cleaved muscovite mica (Asheville-Schoonmaker Mica, Newport News, VA). Suitable regions of the sample with well-isolated molecules were located by surveying the sample using fields-of-view of 2×2 μm^2. Isolated proteins were then imaged using 512×512 nm^2 fields-of-view with 512×512 pixels to yield 1 nm/pixel resolution.

Model regression

Models of the molecules imaged by SFM were obtained using MATLAB version 6.0 (The Mathworks, Natick, MA) on a 666 MHz dual Pentium III PC. Models of each protein were initialized by drawing a set of 10 points along the major axis of the protein backbones to represent initial positions for the cubic spline knots using the GNU Image Manipulation Program (Free Software Foundation, Cambridge, MA). The hoops representing the widths and heights were initially assumed to have dimensions of 1 nm for both the vertical and horizontal diameters.

Image simulation was performed by slightly altering the idilation C function written by Villarrubia (1997) so that it could be called from MATLAB using the ''mex file'' facilities. The quantity χ^2 from Eq. 3 represented the cost function and was used to optimize the fit. A globally convergent nonlinear optimization algorithm (Huyer and Neumaier, 1999) was applied early in the reconstruction process to get into the neighborhood of the (presumed) global minima. Once in the neighborhood, a fast but locally convergent optimization algorithm based on the DIRECT method was used (MATLAB's fminsearch function). The source code for the image reconstruction is available from the authors.

EXPERIMENTAL RESULTS AND DISCUSSION FOR SFM IMAGES OF AGGRECAN

We collected and analyzed 42 high-resolution SFM images of aggrecan. Representative topographs (Fig. 6, *a* and *b*) and examples of isolated aggrecan molecules (Fig. 6, *c–f*) and the corresponding reconstructed protein models (Fig. 6, *g–j*) are shown. Topographs are 512×512 nm^2 with a linear gray-

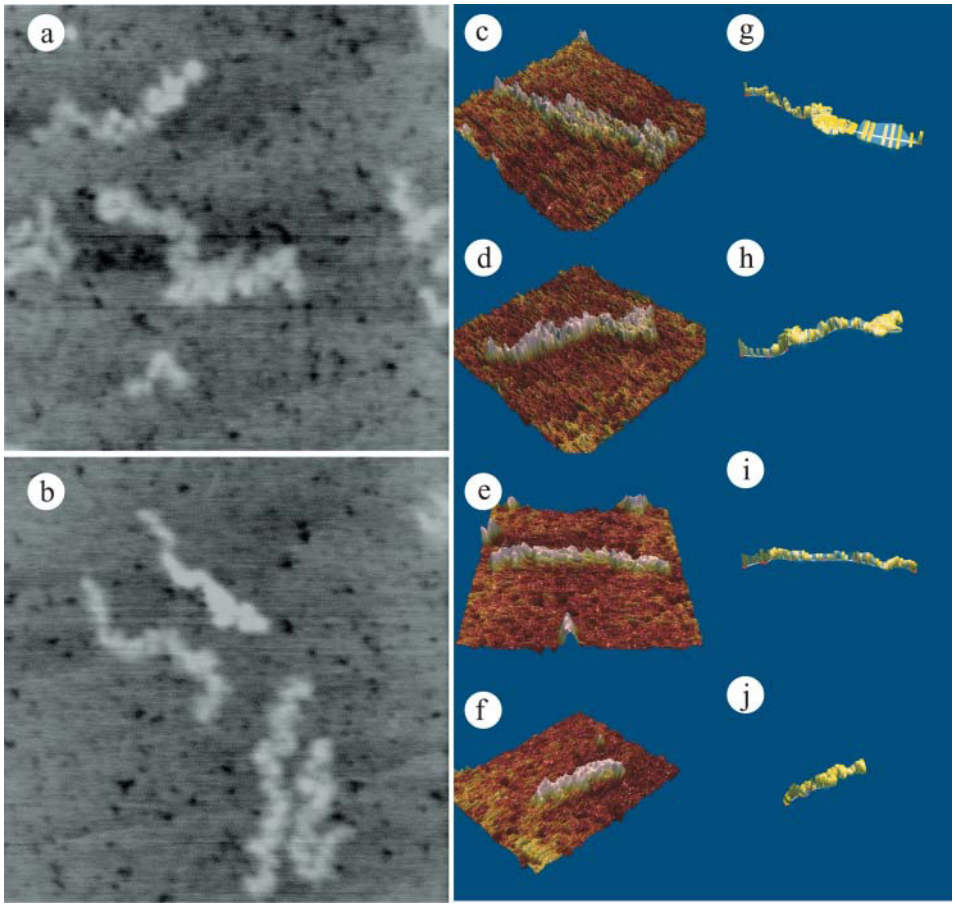

FIGURE 6 Experimental SFM images of aggrecan and corresponding molecular models. (*a* and *b*) Representative pseudocolor topographs, each 512 × 512 nm^2 with a linear gray scale of 1 nm. (*c–f*) Three-dimensional renderings of isolated molecules. (*g–j*) Reconstructed molecular models. Three-dimensional surfaces are 300 nm on a side with the height dimension exaggerated by a factor of 50 to facilitate contrast. All of the images are oriented with the less glycosylated amino-terminal region (*narrow*) of the molecule on the left side. There is a large variation in the size and conformation of molecules.

scale of 1 nm. Three-dimensional renderings are displayed at 300 × 300 nm^2 with the height dimension exaggerated by a factor of 50 to enhance contrast. Many of the molecules exhibited wide and narrow ends. We used this asymmetry to orient the molecules with respect to amino- and carboxyl-termini by assuming that wide ends resulted from the heavily glycosylated carboxyl-terminus and narrow ends from the less glycosylated amino-terminus. Below we show how length, width, and backbone curvature can be quantified from the reconstructed molecular structures. This information is then mapped onto the protein primary sequence to associate three-dimensional structures with biochemically significant regions previously described using cDNA analysis.

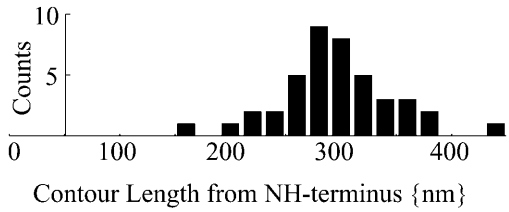

FIGURE 7 The contour length of backbones calculated from the best-fit models of 42 different aggrecan molecules imaged by SFM. The lengths range from 170 to 450 nm with a mean of 295 nm and a standard deviation of 50 nm.

Length

Fig. 7 shows the length distribution of aggrecan molecules. The lengths of molecules ranged from 170 nm to 450 nm with a mean of 295 nm and standard deviation of 50 nm. The lengths are fairly symmetrically distributed about the mean. For comparison, we also made measurements of the lengths using the traditional ''line scan'' analysis of the raw tip broadened image data (distribution not shown). The lengths ranged from 190 to 460 nm with a mean of 310 nm and a standard deviation of 60 nm.

Width

The width of a molecule at a given point along the contour length is the diameter of the horizontal axis of the elliptical cross section at that point. Fig. 8 *a* shows that there tends to be a narrow end (*left*) and a broad end (*right*) of a representative molecule. Fig. 8 *b* plots the width at each contour length position. The transition from narrow to broad

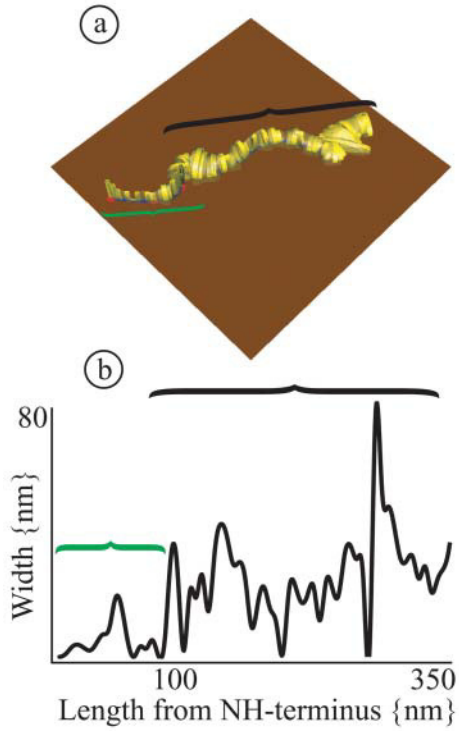

FIGURE 8 The distribution of glycosaminoglycans along the protein backbone are reflected in the widths of the molecular models measured by SFM. (*a*) In the model, we can qualitatively see a distinct narrow end toward the right side and a wide end toward the left side. (*b*) By plotting the width as a function of contour length, we often see a distinct transition around ~100 nm.

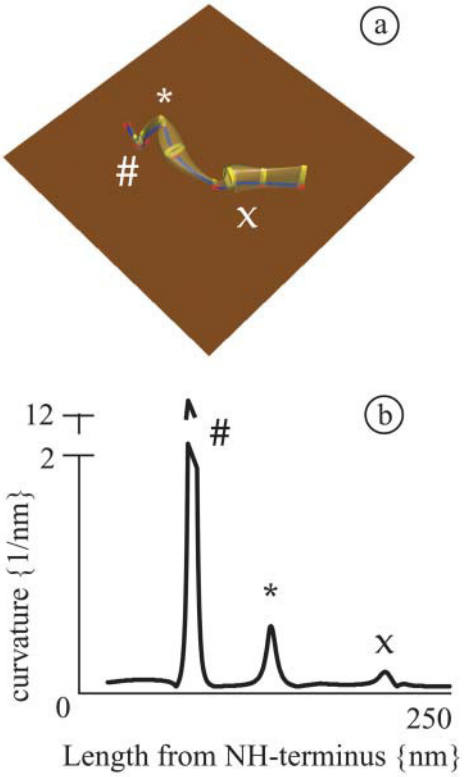

FIGURE 9 The conformation of a protein backbone can be investigated quantitatively by calculating the curvature of the model backbone. (*a*) In the model we see some sharp kinks (indicated by # and *) and some more gradually bending regions (indicated by *x*). (*b*) Plotting the curvature as a function of contour length, the kinks show up as large peaks and the smoothly bending regions as much smaller peaks.

occurs ~100 nm from one end of the molecule. The width data for the entire population of 42 molecules ranged from 0.1 to 72 nm with a mean of 11 nm and standard deviation of 7 nm. A number of molecules exhibited abrupt transitions similar to the molecule in Fig. 8; however, upon averaging the data, no distinct transition was observed.

Molecular curvature

The backbone conformation of a fibrillar molecule imaged by SFM is often of interest because it reflects molecular flexibility. This can be quantified easily using our method because the backbone is represented explicitly in the model. A plot of curvature is shown for one reconstructed protein model (Fig. 9) with annotations that indicate where peaks in the curvature correspond to morphological features in the images. Notice that the image (Fig. 9 *a*) contains some smoothly curving regions and some kinks that are very sharply bent. In the plot of curvature as a function of contour length (Fig. 9 *b*), the kinks appear as large peaks (*y* axis broken to facilitate contrast) and the more smoothly bending regions as relatively small peaks. This information is easily quantified from our models and can be used to determine the flexible regions in a protein backbone (Hofmann et al., 1984) or the persistence length of a polymer (Rivetti et al., 1996, 1998).

Connecting SFM measurements with primary structure locations

Our reconstruction technique parameterizes the three-dimensional structural information in an SFM image along the backbone of a generalized cylindrical model. For example, Fig. 8 shows how the width varies along the backbone and Fig. 9 shows how the curvature varies along the backbone. For fibrillar proteins where the amino acid backbone is relatively extended, there is a strong relationship between this coordinate and amino acid position in the primary sequence. Hence, we can use our modeling technique to relate three-dimensional structural information in the SFM images to the primary structure of the protein, generally determined from cDNA analysis.

To make this connection, we require a mapping between length along the protein backbone and position in the amino acid sequence. Previously, Hofman et al. (1984) performed a similar analysis with collagen where they made a qualitative association between a flexible site observed by TEM and a proline-poor region of type I collagen. Here, we use measurements of domain sizes made previously by TEM (taken from Table 1, Morgelin et al., 1989) and corrected for

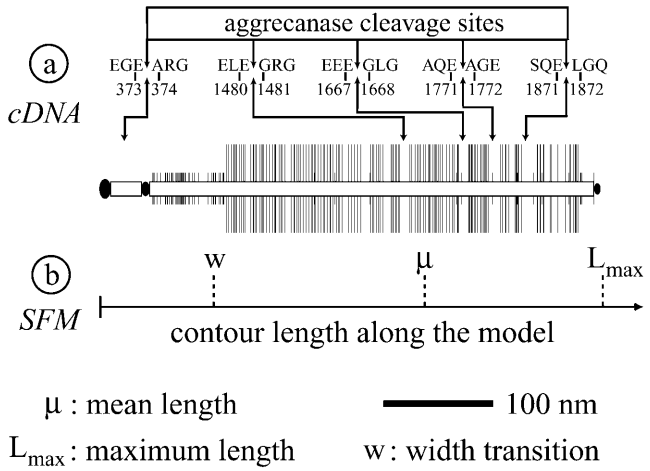

FIGURE 10 Relating SFM measurements of aggrecan to its primary structure derived from cDNA analysis. (*a*) Each domain of the primary sequence shown in Fig. 2 has been scaled according to measurements by TEM (Morgelin et al., 1989) to give a mapping between contour length and amino acid sequence position. See Fig. 2 for more details. (*b*) Width and length measurements from SFM. *w* denotes the position where we observed a transition from a small to large width. The position is very near the start of the chondroitin sulfate attachment domains seen in (*a*). μ and L_{max} denote the mean and maximum length measured by SFM respectively. The mean length lies between two sites in the primary structure known to be cleaved by aggrecanase (Tortorella et al., 1999, 2000), E(1480)-(1481)G and E(1667)-(1668)G; 48% of the molecules we measured had lengths between the expected locations of these cleavage sites.

2.5 nm broadening artifact from the rotary shadowing to obtain a quantitative mapping between contour length and amino acid sequence (Fig. 10 *a*). Note that the globular domains (*black ovals*) decrease in size (going from Fig. 2 to Fig. 10 *a*) relative to the extended domains (*white boxes*). This agrees with secondary structure predictions that the globular domains form compact, disulfide bonded structures, whereas the extended domains form random and β-pleated sheet structures (Hering et al., 1997). Using this mapping, we can take a structural feature (e.g., height, width, or curvature) in an SFM image, find the contour length in Fig. 10 *a*, and associate the feature with a specific location in the cDNA sequence.

Fig. 10 *b* demonstrates this by showing the contour lengths associated with SFM width and length measurements. The *w* in the Fig. 10 *b* denotes where the transition between the narrow and wide region in Fig. 8 occurred (~100 nm from the amino terminus). Comparing this with the cDNA information (Fig. 10 *a*), we see that this corresponds to the position where serine-glycine repeats are found (each site is shown by one of the longer lines running perpendicular to the molecule). These sites are expected to be substituted with chondroitin sulfate chains with average lengths around 36 nm (Morgelin et al., 1989). This shows that our modeling technique detects glycosylation from the width of the model and that by parameterizing width along the contour length, we related the three-dimensional structure to serine-glycine rich domains known from the cDNA sequence (Hering et al., 1997). We also show how the length distribution registers onto the cDNA sequence. The measured 450 nm maximum length (denoted L_{max}) of the molecule agrees with the length of the model based on ''intact'' nondegraded molecules (Morgelin et al., 1989). The mean length of molecules (μ), however, is ~1/3 shorter and agrees with mean lengths of ~300 nm obtained by Buckwalter et al. using TEM (Buckwalter et al., 1985, 1987, 1989; Thonar et al., 1986). Mapping this structural information to the cDNA model, we find the mean length falls between the two kinetically most active aggrecanase cleavage sites at E(1480)-(1481)G and E(1667)-(1668)G (Tortorella et al., 2000); 48% of the molecules we measured had lengths between the expected locations of these cleavage sites.

SUMMARY

We presented a new method for reconstructing biomolecular structures from SFM data. The method works by proposing a model, simulating an image, and using nonlinear regression to adjust the parameters to best-fit the experimental data. Under the assumption that the model is capable of representing the actual structure of the molecule, this technique can accurately recover molecular dimensions from tip-broadened SFM data. In contrast, previous methods relying on reconstruction by erosion (Eppell et al., 1993; Keller, 1991; Markiewicz and Goh, 1994; Wilson et al., 1996) are capable only of placing an upper bound on the dimensions of a molecular structure (Villarrubia, 1997). It should be stressed, however, that the solutions obtained from our technique are nonunique; other models may fit the data as well or better and there is no way to tell from the SFM images alone which best represents the physical structure of the molecule. Nevertheless, structures imaged by SFM are almost invariably interpreted in terms of some model when subsets of ''characteristic dimensions'' are extracted from the image by taking cross sections (so-called ''line scans''). The technique that we apply here substantially improves this process by including the model explicitly, accounting for the dominant instrument artifact, and converging to a statistically significant structure that includes all of the available data and not some small subset inspected by a line scan.

We demonstrated the facility of the technique by reconstructing molecular models from SFM images of the cartilage proteoglycan aggrecan. The mean length that we measured using the method was 5% smaller than that measured from the raw SFM data with 20% less variability (standard deviation of 50 nm compared to 60 nm). Furthermore, the mean length obtained using the new method was in close agreement with measurements made previously by TEM (Buckwalter et al., 1985, 1987, 1989; Thonar et al., 1986). This supports the results of our simulation example

that showed that the method reconstructs lateral dimensions consistent with TEM.

The generalized cylinder model that we used to describe the shape of aggrecan facilitated quantification of three-dimensional structures including, length, width, and backbone curvature. Extracting this information from SFM images is an otherwise nontrivial problem (Rivetti and Codeluppi, 2001; Rivetti et al., 1998). By using contour length as the independent variable in the model, we were able to relate tertiary structure measured by SFM with the primary structure of the molecule known from cDNA (Hering et al., 1997). This is a luxury afforded by fibrillar proteins because of the simple relationship between conformation and sequence. However, it demonstrates that modeling of molecular structure is an effective means to connect SFM images with data of a very different nature.

This connection allowed us to associate structures observed in the SFM data with primary structural locations known to be important in the degradation of aggrecan in cartilage (Tortorella et al., 1999, 2000). Our results suggest that a large fraction of aggrecan molecules in vivo (48% by our data) have been cleaved at E(1480)-(1481)G and/or E(1667)-(1668)G. This compares favorably with biochemical data that showed 30–50% of aggrecan molecules in vivo lack the carboxyl-terminal G3 domain (Flannery et al., 1992) and in vitro biochemical studies that showed E(1480)-(1481)G and E(1667)-(1668)G to be the most kinetically favorable of the five aggrecanase cleavage sites. This ability to associate three-dimensional structure with primary structure (genetic) information suggests that SFM has the potential for proteomic investigations that cannot easily be studied by XRD or NMR.

We thank Dr. Jiann-Jiu (James) Wu for providing types II, IX, and XI collagen. Aggrecan was generously provided by Prof. Thomas Hering's laboratory.

We also thank the Whitaker Foundation and the National Institutes of Health, grant No. AR45664-01, for generous financial support.

REFERENCES

Abola, E., P. Kuhn, T. Earnest, and R. C. Stevens. 2000. Automation of X-ray crystallography. *Nat. Struct. Biol.* 7 (Suppl):973–977.

Bella, J., M. Eaton, B. Brodsky, and H. M. Berman. 1994. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science*. 266:75–81.

Binnig, G., C. F. Quate, and C. Gerber. 1986. Atomic force microscope. *Phys. Rev. Lett.* 56:930–933.

Block, J. A., S. E. Inerot, and J. H. Kimura. 1992. Heterogeneity of keratan sulfate substituted on human chondrocytic large proteoglycans. *J. Biol. Chem.* 267:7245–7252.

Buckwalter, J. A., K. E. Kuettner, and E. J. Thonar. 1985. Age-related changes in articular cartilage proteoglycans: electron microscopic studies. *J. Orthop. Res.* 3:251–257.

Buckwalter, J. A., L. C. Rosenberg, and R. Ungar. 1987. Changes in proteoglycan aggregates during cartilage mineralization. *Calcif. Tissue Int.* 41:228–236.

Buckwalter, J. A., K. C. Smith, L. E. Kazarien, L. C. Rosenberg, and R. Ungar. 1989. Articular cartilage and intervertebral disc proteoglycans differ in structure: an electron microscopic study. *J. Orthop. Res.* 7:146–151.

Czajkowsky, D. M., and Z. Shao. 1998. Submolecular resolution of single macromolecules with atomic force microscopy. *FEBS Lett.* 430:51–54.

Eppell, S. J., F. R. Zypman, and R. E. Marchant. 1993. Probing the resolution limits and tip interactions of atomic-force microscopy in the study of globular proteins. *Langmuir*. 9:2281–2288.

Ernst, F., and M. Ruhle. 1997. Present developments in high-resolution transmission electron microscopy. *Current Opinion in Solid State & Materials Science*. 2:469–476.

Flannery, C., V. Stanescu, M. Morgelin, R. Boynton, J. Gordy, and J. Sandy. 1992. Variability in the G3 domain content of bovine aggrecan from cartilage extracts and chondrocyte cultures. *Arch. Biochem. Biophys.* 297:52–60.

Harris, J. W., and H. Stocker. 1998. Generalized Cylinder. Handbook of Mathematics and Computational Science. Springer-Verlag, New York.

Hascall, V. C., and S. W. Sajdera. 1970. Physical properties and polydispersity of proteoglycan from bovine nasal cartilage. *J. Biol. Chem.* 245:4920–4930.

Heinegard, D., and I. Axelsson. 1977. Distribution of keratan sulfate in cartilage proteoglycans. *J. Biol. Chem.* 252:1971–1979.

Hering, T. M., J. Kollar, and T. D. Huynh. 1997. Complete coding sequence of bovine aggrecan: comparative structural analysis. *Arch. Biochem. Biophys.* 345:259–270.

Hofmann, H., T. Voss, K. Kuhn, and J. Engel. 1984. Localization of flexible sites in thread-like molecules from electron micrographs. Comparison of interstitial, basement membrane and intima collagens. *J. Mol. Biol.* 172:325–343.

Huyer, W., and A. Neumaier. 1999. Global optimization by multilevel coordinate search. *Journal of Global Optimization*. 14:331–355.

Keller, D. 1991. Reconstruction of Stm and Afm images distorted by finite-size tips. *Surface Science*. 253:353–364.

Lamzin, V. S., and A. Perrakis. 2000. Current state of automated crystallographic data analysis. *Nat. Struct. Biol.* 7 (Suppl):978–981.

Markiewicz, P., and M. C. Goh. 1994. Atomic-force microscopy probe tip visualization and improvement of images using a simple deconvolution procedure. *Langmuir*. 10:5–7.

Montelione, G. T., D. Zheng, Y. J. Huang, K. C. Gunsalus, and T. Szyperski. 2000. Protein NMR spectroscopy in structural genomics. *Nat. Struct. Biol.* 7 (Suppl):982–985.

Morgelin, M., M. Paulsson, A. Malmstrom, and D. Heinegard. 1989. Shared and distinct structural features of interstitial proteoglycans from different bovine tissues revealed by electron microscopy. *J. Biol. Chem.* 264:12080–12090.

Rivetti, C., and S. Codeluppi. 2001. Accurate length determination of DNA molecules visualized by atomic force microscopy: evidence for a partial B- to A-form transition on mica. *Ultramicroscopy*. 87:55–66.

Rivetti, C., M. Guthold, and C. Bustamante. 1996. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J. Mol. Biol.* 264:919–932.

Rivetti, C., C. Walker, and C. Bustamante. 1998. Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J. Mol. Biol.* 280:41–59.

Rosenberg, L., H. U. Choi, L. H. Tang, S. Pal, T. Johnson, D. A. Lyons, and T. M. Laue. 1991. Proteoglycans of bovine articular cartilage. The effects of divalent cations on the biochemical properties of link protein. *J. Biol. Chem.* 266:7016–7024.

Thonar, E. J., J. A. Buckwalter, and K. E. Kuettner. 1986. Maturation-related differences in the structure and composition of proteoglycans synthesized by chondrocytes from bovine articular cartilage. *J. Biol. Chem.* 261:2467–2474.

Todd, B. A., and S. J. Eppell. 2001. A method to improve the quantitative analysis of SFM images at the nanoscale. *Surface Science*. 491:473–483.

Tortorella, M. D., T. C. Burn, M. A. Pratta, I. Abbaszade, J. M. Hollis, R. Liu, S. A. Rosenfeld, R. A. Copeland, C. P. Decicco, R. Wynn, A. Rockwell, F. Yang, J. L. Duke, K. Solomon, H. George, R. Bruckner, H. Nagase, Y. Itoh, D. M. Ellis, H. Ross, B. H. Wiswall, K. Murphy, M. C. Hillman, Jr., G. F. Hollis, E. C. Arner, et al. 1999. Purification and cloning of aggrecanase-1: a member of the ADAMTS family of proteins. *Science*. 284:1664–1666.

Tortorella, M. D., M. Pratta, R. Q. Liu, J. Austin, O. H. Ross, I. Abbaszade, T. Burn, and E. Arner. 2000. Sites of aggrecan cleavage by recombinant human aggrecanase-1 (ADAMTS-4). *J. Biol. Chem.* 275:18566–18573.

Villarrubia, J. 1997. Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation. *J. Res. Natl. Inst. Stand. Technol.* 102:425–454.

Wilson, D. L., P. Dalal, K. S. Kump, W. Benard, P. Xue, R. E. Marchant, and S. J. Eppell. 1996. Morphological modeling of atomic force microscopy imaging including nanostructure probes and fibrinogen molecules. *Journal of Vacuum Science & Technology B*. 14:2407–2416.

Wilson, D. L., K. S. Kump, S. J. Eppell, and R. E. Marchant. 1995. Morphological restoration of atomic force microscopy images. *Langmuir*. 11:265–272.