Strategy for far-field optical imaging and writing without diffraction limit

Stefan W. Hell

Max-Planck-Institute for Biophysical Chemistry, Department of NanoBiophotonics, Göttingen, Germany

Received 28 January 2004; received in revised form 22 March 2004; accepted 22 March 2004

Communicated by B. Fricke

Abstract

A general concept is described for subdiffraction imaging and writing with focused light and conventional optics. It is unique in the fact that the giant optical nonlinearities required for breaking the diffraction barrier are realized at low light intensities. The result is that far-field resolution on the nanoscale becomes feasible at realistic physical conditions. The predicted subdiffraction resolution or feature size is formulated in a simple equation.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Optical microscopy; Optical structuring; Nonlinear optics; Nanoscale resolution; Diffraction barrier; Saturable molecular transitions

1. Introduction

It is common knowledge that diffraction prevents focused light from imaging and defining spatial structures much smaller than the wavelength. Diffraction has so far precluded the detailed visualization of viable cellular organelles and, by the same token, the writing of nanostructures with visible focused light. Therefore, large technological efforts are being undertaken to push data storage to shorter wavelengths and lithography into the X-ray regime.

However, in the last few years concepts have been proposed that detailed for the first time how to break the diffraction barrier in far-field fluorescence microscopy [1,2]. Having the potential to provide diffraction-unlimited resolution, these concepts took a fundamental step beyond the established superresolution techniques. Stimulated emission depletion (STED) microscopy, for example, has so far displayed a spatial resolution of ~30 nm corresponding to $\lambda/25$ at 750 nm [3]. The second concept of this type, ground state depletion microscopy has a similar resolution potential [2]. If we leave aside the technical aspects of each of these concepts, a common principle comes into view: the diffraction barrier is overcome by the saturation of a reversible molecular transition with an intensity distribution featuring a strong spatial gradient. In this Letter, after generalising the underlying principle, we outline how it can be extended to writing spatial features of arbitrary width and density.
It has for long been clear that the breaking of the diffraction barrier requires optical nonlinearities. As large nonlinearities are often connected with strong light fields, the prevailing notion has been that nanoscale resolution would be precluded by the requirement of unacceptably high intensities. Here we show that the ‘intensity barrier’ to attaining nanoscale far-field resolution can be outperformed as well. Pieced together, these aspects allow us to propose a strategy for mastering the nanoscale with focused light and conventional optics. Concrete proposals for experimental implementations are given.

2. Basic concept

Let us assume a light absorbing entity (e.g., a molecule, cluster, solid) that can be transferred between two states A and B at any point in space. In the simplest case, A and B are ordinary energy states of this entity, such as a ground and an excited state. Other examples are molecular conformational states, photochromic and isomerization states, binding and protonation states, etc. As we wish to image or write with light, the transition A → B is photoinducible. We make no specific restrictions about B → A. In the most general case, B → A will have a spontaneous component together with a component which is triggered externally through light, heat, by a chemical reaction, etc. For simplicity, the substeps of A → B and B → A should be different, i.e., the entity assumes different intermediate states during the cycle (Fig. 1), making coherent effects negligible.

By denoting the rates of A → B and B → A with \( k_{AB} \) and \( k_{BA} \), respectively, the normalized populations \( N_A \) and \( N_B \) are described by

\[
\frac{dN_A}{dt} = -k_{AB}N_A + k_{BA}N_B = -\frac{dN_B}{dt}.
\]

If it first resides in A, the probability of still finding the entity in A after time \( t \) is:

\[
N_A(t) = \frac{k_{BA}}{k_{AB} + k_{BA}} \left( \frac{k_{BA}}{k_{AB}} + \exp\left[-(k_{AB} + k_{BA})t\right] \right).
\]  

(1)

At \( t > 5(k_{AB} + k_{BA})^{-1} \), the equilibrium \( N_A^{\infty} \) is reached within \(< 5\%\).

We are now interested in shifting the entity to B. The transition A → B should be affected at a wavelength \( \lambda_{AB} \), with a rate \( \sigma_{AB}I_{AB} \), whereby \( \sigma_{AB} \) and \( I_{AB} \) denote the molecular cross-section and the local photon flux per unit area, respectively. To be as general as possible, we assume that \( I_{AB} \) may also undesirably induce B → A with a cross-section \( \sigma_{BA}^{ab} \). Therefore, in the most general case, B → A consists of the externally triggered component \( E \), plus a spontaneous component \( k_{BA}^s \), so that \( k_{BA} = E + k_{BA}^s + \sigma_{BA}^{ab} I_{AB} \); see Fig. 1. Hence, it follows

\[
N_A^{\infty} = \frac{\sigma_{BA}^{ab} I_{AB} + E + k_{BA}^s}{(\sigma_{AB} + \sigma_{BA}^{ab})I_{AB} + E + k_{BA}^s}.
\]  

(2)

For intensities

\[
I_{AB} \gg \frac{E + k_{BA}^s}{\sigma_{AB} - \sigma_{BA}^{ab}} \equiv I_{AB}^{sat},
\]

we obtain \( \bar{N}_A^{\infty} = \sigma_{BA}^{ab} I_{AB}^{sat} / (\sigma_{AB} + \sigma_{BA}^{ab}) \). To attain a marked shift to B, i.e., \( \bar{N}_A^{\infty} \approx 0 \), one obviously prefers \( E = 0 \) and adjusts \( \lambda_{AB} \) so that \( \sigma_{BA}^{ab} \ll \sigma_{AB} \).

Let A → B be now caused by a spatial intensity distribution featuring a local zero: \( I_{AB}^{(r)}(\vec{r}) = I_{AB}^{(r_{0})} f(\vec{r}) \), with \( f(\vec{r}_{0}) = 0 \). \( f(\vec{r} = x, y, z) \) is a normalized function describing the diffraction-limited distribution of the intensity. If we now elect \( I_{AB}^{\text{max}} \gg I_{AB}^{\text{sat}} \), we obtain \( \bar{N}_A^{\infty} = \bar{N}_A^{\infty} \) at every point in space, except at the closest proximity of \( \vec{r}_{0} \) where \( \bar{N}_A^{\infty} = 1 \). Thus we can create an arbitrarily sharp region that is purely in A, which is
embedded into a region with a marked B component. Importantly, the arbitrarily sharp region of state A neither depends on $\lambda_{AB}$ nor on the details of $f(\vec{r})$. It is irrelevant whether $f(\vec{r})$ describes a doughnut-mode or a standing wave, or any other function if only it features at least a single local zero.

In most practical situations the $I_{AB}^\text{max}$ and $I_{AB}^\text{sat}$ are finite and $f(\vec{r})$ is a distinct function. The perhaps most simple function is a standing wave $f(x) = \sin^2(2\pi x/\lambda_{AB})$, in which case the full-width-half-maximum of the narrow ‘A state region’ can readily be calculated:

$$\Delta x \geq \frac{\lambda_{AB}}{\pi} \arcsin \left( \frac{E + k_{BA}^2}{(\sigma_{AB} - \sigma_{BA}) I_{AB}^\text{max}} \right)$$

$$= \frac{\lambda_{AB}}{\pi} \arcsin(\varsigma^{-1/2}) \approx \frac{\lambda_{AB}}{\pi \sqrt{\varsigma}}. \quad (3)$$

The latter approximation obviously holds for saturation factors $\varsigma = I_{AB}^\text{max} / I_{AB}^\text{sat} \gg 1$. A factor $\varsigma = 1000$ yields $\Delta x \approx \lambda_{AB}/100$, but in principle the spot can be squeezed down to the molecular scale.

Imaging is now performed by scanning an intensity naught across the sample and detecting the molecules in state A. By scanning, all molecules are consecutively transferred to B, except for those located within the $\Delta x$ region around the naught. Consequently, sub-diffraction sized sample features can be mapped out through the molecules that have been (transiently) left in A. If the state A is a fluorescent state, the image is gained by sequentially storing the recorded fluorescence from the spatial interval $\Delta x$. Since the image is assembled point by point and the signal is integrated in a detector, the diffraction-limited bandwidth of the optical components is irrelevant. The diffraction limited bandwidth only comes into play if several nodes are read out in parallel by imaging them onto a detector array or camera. In this case, one has to make sure that the nodes are further apart than the diffraction resolution limit of the imaging system, so that the spot images are spatially distinct at the detector array. If the initial spatial distribution of state A follows $g(x) = \cos^2(2\pi x/\lambda_{AB})$, we obtain

$$\Delta x \geq \frac{\lambda_{AB}}{\pi} \arcsin \left( \frac{1}{\sqrt{\varsigma + 2}} \right) \approx \frac{\lambda_{AB}}{\pi \sqrt{\varsigma + 2}}. \quad (4)$$

For $I_{AB}^\text{max}$, $\varsigma \to 0$, we have $\Delta x = \lambda_{AB}/4$ as expected for a standing wave. If $f(x)$ and $g(x)$ are produced by a lens of numerical aperture $n \sin \alpha$, we have $\lambda_{AB} = \lambda/(n \sin \alpha)$ as the shortest effective wavelength possible and (4) turns into

$$\Delta x \geq \frac{\lambda}{\pi n \sin \alpha} \arcsin \left( \frac{1}{\sqrt{\varsigma + 2}} \right)$$

$$\approx \frac{\lambda}{\pi n \sin \alpha \sqrt{\varsigma + 2}}, \quad (5)$$

which for $\varsigma \to 0$ delivers the peak width corresponding to a $\pi$ phase shift of Abbe’s ultimate wave period $\lambda/(2n \sin \alpha)$. Again, for $\varsigma \to \infty$ the spatial resolution of the scanning imaging system becomes infinite.

3. Writing at the nanoscale

If the utilized material can be transferred from the state A into a permanent state C, a new concept can be defined for writing structures without diffraction limit. First, the material residing in state A is transferred to the non-writable state B by $I_{AB}(\vec{r}) = I_{AB}^\text{max} f(x)$, leaving at its minima $x_i$ writable structures of A of dimension $\Delta x \ll \lambda_{AB}$. Next, this structure is made permanent by initiating the transition A $\to$ C. To repeat the process, the intensity pattern is shifted to $x_i + \delta x$, with $\delta x \geq \Delta x/2$. The procedure is illustrated in Fig. 2. Whereas the role of the saturable transition A $\to$ B is to define the location and width of the structure, the role of B is to protect the material from being written. Hence, the fundamental difference to established writing processes is the existence of a transient protective state B that is inert to the writing procedure.

If the transition B $\to$ A can be induced by light, thermally, chemically, or by any other means (i.e., the rate $E$), the timing of the process can be fully controlled. Importantly, both the width $\Delta x$ and the distance $\delta x$ between the structures are arbitrarily scalable. As a result, the concept allows the free writing of structures of any dimension and density with focused light, irrespective of the wavelength used.

4. Strong nonlinearities at low intensities

Since Eq. (3) does not present fundamental limits on the attainable resolution, the real limit will be
set by the saturation level supported by the material. Therefore, the strategy of choice is to choose saturable transitions featuring a low saturation intensity $I_{\text{sat}}^{\text{AB}}$.

In STED-fluorescence microscopy, which is now viewed as a special case of this concept, the marker fluorophores in the fluorescent state $S_1$ (state A) are stimulated to the ground state $S_0$ (B) in a saturated manner. The excitation of the dye (B $\to$ A), described by the rate $E$, completes the cycle. The substeps of the cycle occur via different vibrational sublevels, so that the two transitions are distinct. Organic fluorophores feature $\sigma_{\text{AB}} \approx 10^{-16}$ cm$^2 \gg \sigma_{\text{BA}}^{\text{sat}}$ and a stable ground state, but since A spontaneously decays at $k_{\text{AB}}^{\text{sat}} \approx 10^9$ s$^{-1}$, we have

$$I_{\text{sat}}^{\text{AB}} = \frac{k_B^{\text{sat}}}{\sigma_{\text{AB}}} \approx 10^{25} \frac{\text{photons}}{\text{s cm}^2} \approx 1-5 \frac{\text{MW}}{\text{cm}^2}.$$  

For $I_{\text{AB}}^{\text{max}} \gg I_{\text{AB}}^{\text{sat}}$, the stimulated transition A $\to$ B is predominant. So far, factors $\xi \approx 120$ have been experimentally reported [4], showing that the potential of STED to overcome the diffraction barrier is in the region of a factor of 10. The realization of an even larger increase in resolution will require $\xi > 10^3$ and therefore might involve intensities that are not tolerable in fluorescence microscopy. The alternative to deplete the ground state $S_0$ (now A) by saturating the excited state $S_1$ (now B) [2,5], does not present a solution to the intensity problem. In this case, $I_{\text{AB}}^{\text{sat}}$ is of the same order as in STED [3], because $\sigma_{\text{AB}}$ is similar in magnitude and the $S_1$ decays at the same rate $k_{\text{AB}}^{\text{sat}}$. Therefore, the obvious way to reduce $I_{\text{AB}}^{\text{sat}}$ is to target an excited state (B) with a longer lifetime compared to $S_1$, such as a metastable state.

The potential of this strategy is illustrated with several examples. A first example is the metastable triplet state $T_1^0$. Due to the strong intersystem crossing, in many fluorophores the state $T_1^0$ can be reached through the $S_1$ with effective cross sections $\kappa_{\text{AB}}$: values of $10^{-2} < \kappa < 10^{-1}$ are typical [6]. Since the molecule relaxes to A with a rate $10^1$ s$^{-1} < k_{\text{BA}}^s < 10^6$ s$^{-1}$, pushing it to $T_1^0$ entails $10^2–10^5$ times lower $I_{\text{AB}}^{\text{sat}} = 10^{–2}–10$ kW/cm$^2$. The signal from $r_0$ to be measured is the fluorescence of the molecules that remained in the singlet system.

Similarly, low $I_{\text{AB}}^{\text{sat}}$ is attained by filling up the metastable states of rare earth metal ions, that are energy-fed through light-absorbing chelates [6]. Since the chelates feature regular absorption cross-sections of the order of $10^{-16}$ cm$^2$ and the energy transfer is in the range of 50–100%, the $\sim 1$ ms lifetime of the metastable state enables saturation intensities $I_{\text{AB}}^{\text{sat}}$ that are lower by 5–6 orders of magnitude than with STED.

It becomes clear that the lowest possible saturation intensities $I_{\text{AB}}^{\text{sat}}$ are attained if the spontaneous rates $k_{\text{AB}}^s$ and $k_{\text{BA}}^s$ are vanishing. In other words, the states A and B are bistable and the molecule can be switched back and forth. As there is no counteracting transition, the application of a finite intensity already causes $\Delta x$ to shrink without limit, with its size depending only on the exposure time. The only additional issue to be considered is that the molecule has to be actively brought back to A, for example, by effecting another absorption of light at a different wavelength. Thus
the utilization of compounds with two (semi-)stable states abolishes the high intensity issue altogether. This is a crucial insight since it makes evident that extremely high saturation factors and hence huge optical nonlinearities can be achieved at very low focal intensities.

Concrete examples of suitable semi-stable compounds are photoswitchable fluorescent proteins and bistable photochromic materials. For example, the recently reported fluorescent protein aSFP595 [7] should require switching intensities that are below 1 W/cm². Switchable marker proteins thus have the potential to provide huge optical nonlinearities through saturation. As a result, spatial resolution of < 5 nm should be achievable in live cells without intensity hazards. The advantage of using fluorescent proteins is that they are genetically encodable. Moreover, synthetic photochromic compounds, such as fulgides, diarylethenes, etc., that are currently being developed for optical data storage should also be similarly suitable. For example, a reversibly photoswitchable fluorescent dye system employing a diarylethene derivative has recently been reported [8]. This molecular system can be transferred between a fluorescent state A and a non-fluorescent state B that are both thermally stable. The switch occurs at ‘off’ and ‘on’ wavelengths centered in the UV and visible range, respectively. The data in the literature allows one to predict that Δx < 5 nm should be attained by focusing less than 100 µW of the ‘switch-off light’ to an area of 10⁻⁸ cm² for 50 µs. These examples underscore that the concept can be implemented at benign intensities. Moreover, by exploiting specific properties of suitable compounds, the physical obstacles to image with focused visible light at the nanoscale can be overcome under realistic physical conditions.

5. Discussion

The conditions for implementing this concept are more relaxed than in the model case above. For example, complete depletion of A (N_A∞ = 0) in the non-zero region is not required. It is sufficient that the non-zero region features a large enough population B, so that it can be distinguished from the sharp region of state A. This applies both to imaging and writing. In imaging, N_A∞ ≠ 0 leads to a conventional background image which can be subtracted, provided that the signal to noise ratio is acceptable. In writing, the process can be repeated at the same spot until sufficient contrast is reached. This is readily understood on the basis that the diffraction barrier is overcome by the gradient induced by the nonlinearity of the saturation, which is largely independent of the actual level N_A∞.

In fluorescence imaging, it is not mandatory that A is the fluorescence state. Since the populations A and B are complementary, the subdiffraction image is encoded in both. Reading out state B does not directly yield the superresolved image, but under suitable signal to noise conditions, the superresolved information can be extracted through image processing. Furthermore, the concept is not limited to the use of a single zero-point in space. One can readily parallelize the scanning procedure by utilizing arrays of spots, provided that the zeros are further apart than the diffraction-limited resolving power of the employed objective lens of ≈ λ/2. In this case, one can still associate the diffraction limited spots on the image plane with each individual zero point so that camera based detection becomes possible. In fact, low intensities facilitate parallelization with an array of minima, which might even be mandatory in imaging applications with moving objects. In the end, nanoscale imaging will only be challenged by (Brownian) motion, unless the flux of detected photons is adequate.

The fundamental progress brought about by this concept is that huge optical nonlinearities are generated at low intensities. This stems from the fact that the nonlinearity created by saturation originates from the population kinetics of the states, rather than from the cooperative action of multiple photons. The latter inevitably rely on high intensities, which is the major reason why no multiphoton approach has yet been identified to provide nanoscale resolution in the far-field. By defining k_A B = σ_A B(I_A B)n, the study is readily extended to n-photon absorption, which further sharpens the saturation curve. However, this benefit is unlikely to outweigh the benefits from a low intensity single photon transition.

It is interesting to note that the concept can also be extended to reflective, ferroelectric, or scattering substrates, such as metal films, colloids or semiconductors, that become (semi-) permanently or transiently transparent, or change their spectral properties upon absorption saturation. In complete analogy, the reflec-
tive area shrinks to a minimum defined by the non-saturated regions.

6. Conclusion

Saturable optical transition featuring a reverse transition allow the definition of a range of diffraction-unlimited imaging modes. With suitable materials, the concept can be extended to diffraction-unlimited writing. The resolution depends on the saturation level and can be readily expressed in a simple equation. While the viability of this approach has been demonstrated with saturated stimulated emission for imaging, the generalization paves the way for a new range of diffraction-unlimited imaging and writing techniques that belong to a common family. The need for high intensities that appeared prohibitive for achieving molecular scale resolution can be overcome by involving states with lifetimes $> 1 \mu s$. While a fair number of the presently available compounds have been identified, the strategic synthesis of molecules with reversible saturable transitions offers a unique way to open up the nanoscale with visible focused light. The strategy presented here should spurn research activities to exploit this hitherto unrecognized opportunity.

References