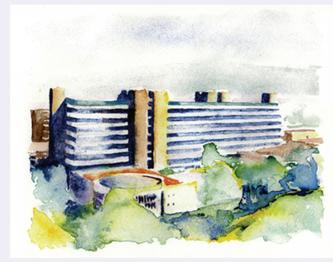
Approaching the Sequential and Three - Dimensional Organization of Archaea, Bacteria and Eukarya Genomes

Tobias A. Knoch [1], Markus Göker [2], Rudolf Lohner [3] and Jörg Langowski [1]

[1] Division Biophysics of Macromolecules, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280
D - 69120 Heidelberg, Federal Republic of Germany
http://www.DKFZ-Heidelberg.de/Macromol/Welcome.html

2) Institute for Special Botanics and Mycology, University of Tübingen,
Auf der Morgenstelle 1, 72076 Tübingen, FRG

3) Computing Center, University of Karlsruhe
76128 Karlsruhe, FRG



THEORY

The analyses of the sequential organization of genomes attempted here is based on the root mean concentration fluctuation function C(l) of the concentration profile of purines/pyrimidines along the DNA sequence: from the base concentration in a window of l bases, the base concentration in the whole base sequence of length L is subtracted and squared. The average is taken over all s=L-l+1 windows and finally the square root is computed:

$$C(l) = \sqrt{\langle (c_l - \bar{c}_L)^2 \rangle_s}$$

To avoid numeric instability (Fig. 1A&B) the equation is expanded to

$$C(l) = \frac{1}{Ll}\sqrt{\frac{1}{L-l}\sum_{s=1}^{L-l}\left[\left(\sum_{k=1}^{l} Ln\right) - \left(\sum_{k=1}^{L} lN\right)\right]^2}$$

with n,N being the base probability e. g. P = 1 for purines and else 0. Since C(l) follows a power-law behaviour, the slope is the correlation coefficient δ.
Beyond non-random simple power-law correlations almost on the entire length scale of sequences (Fig. 1A), δ varies locally on different scales, indicating so called multi-scaling behaviour (Fig. 1A&C). The correlations deviate also clearly from those of random sequences up to large scales (Fig. 1A&C, Fig. 2A&B) before the cut-off due to the finite sequence length. The information relevant for tree construction is located at either small or large scales (Fig. 2C).

Fig. 1: Root mean squared concentration fluctuation function C(l) (A) and its slope, the correlation coefficient δ(l) (B), of random sequences and *Homo sapiens*. The error made by using unstable numerics (A,red; B).

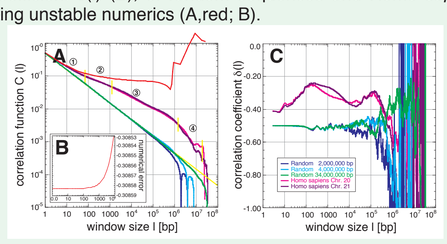
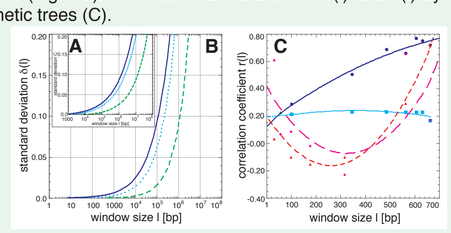


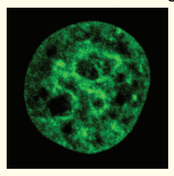
Fig. 2: Standard deviation of C(l) (A) and δ(l) (B) of the fluctuations appearing in random sequences (Fig. 1C). Information distribution in C(l) and δ(l) by correlating trees to phylogenetic trees (C).



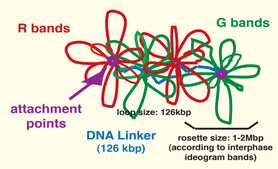
INTRODUCTION

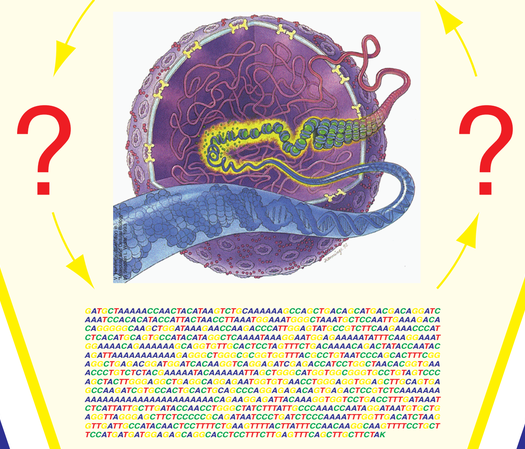
Despite the successful linear sequencing of the genome its sequential organization, i. e. the relations within sequences, and its connection to the three-dimensional structure is widely unknown despite its importance for gene regulation and replication. Through correlation analyses of completely sequenced Archaea, Bacteria and Eukarya genomes, we show an approach relating the sequential to the three-dimensional genome organization.

Nuclear chromatin morphology by histone H2A-YFP *in vivo* labelling



Multi-Loop-Subcompartment (MLS) model of chromatin folding (Münkel *et al.*, 1997)





SIMULATION

The creation of arteficial random sequences using blocks (Fig. 9), codon-triplets (Fig. 10) and a nucleosome binding consensus sequence (Fig. 11), using random or gene-like block arrangements (the latter two) as well as different base pair or codon/block concentrations, explain the multi-scaling and fine-structured long-range correlations of Eukarya (Fig. 3-5), Archaea and Bacteria (Fig. 13). The multi-scaling behaviour is due to blocks and the fine-structure due to the codon usuage and and in *Homo sapiens* nucleosome binding. The periodicity of 12 bp for *Pseudomonas aeruginosa* PA01 cannot be reproduced by pure codon-usage and might be due to a special syntax of codons.

Fig. 9: Simulation of random seqences with different sized (A) and concentrated (B) blocks.

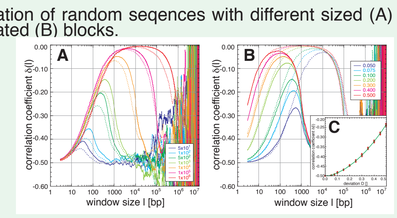


Fig. 10: Simulation and comparison to real sequences of random sequences with random codon triplets, distributed according to the real codon usage-tabels. The codons were either mixed under a random sequence or arranged in gene-like blocks, using different concentrations.

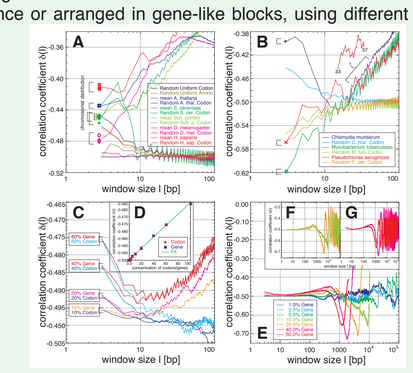
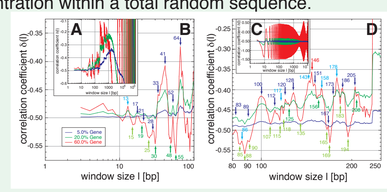


Fig. 12: Simulation of random sequences using a nucleosome binding consensus sequence arranged in gene-like blocks and different block concentration within a total random sequence.

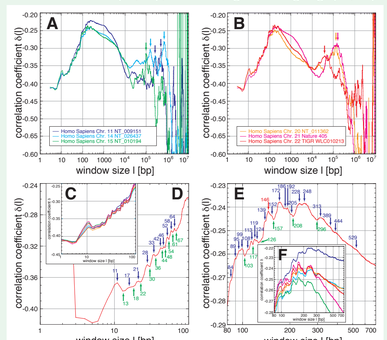


CONCLUSION

Correlation analysis revealed the sequential organiza-tion, and its connection to the three-dimensional organization of genomes. All sequences showed long-range power-law correlations on almost the entire observable scale with different local correlation coefficients. This multi-scaling behaviour was species specific and is due to a block organization of geno-mes as shown by sequence simulations. A fine-struc-ture is present and attributable to the codon usage and to nucleosomal binding. Computer generated random sequences assuming the codon usage and nucleosomal binding agree with these results. Muta-tion by sequence reshuffling destroyed all correla-tions, thus their stability seems evolutionary tightly controlled and connected to the spatial genome organization on large scales. Trees constructed from the correlation behaviour were as expected for Eukarya and led to a new classification system for Archaea and Bacteria. In summary, these findings suggest a complex sequential organization of genomes closely connected to their three-dimensional organization.

EUKARYA GENOMES

Fig. 3: *Homo sapiens* multi-scaling (A, B) and fine-structure (C, F), surviving averaging (D, E).



All analysed Eukarya genomes show beyond a simple long-range power-law, different local correlation coefficients: A first region with one (Fig. 3, 5) or two (Fig. 4) maxima between 50 and 2000 bp is followed by a region of one or more second maxima at ~10^5 bp. This multi-scaling behaviour is species specific and can be explained by a block organization of genomes. Within the multi-scaling a fine-structure is present which survives averaging. It is attributable to the codon usage for lower Eukarya (Fig. 4, 5) and solely to nucleosomal binding in the case of *Homo sapiens* (Fig. 3). The second maxima region and fine-structure of *Homo sapiens* might be due to a folding of the chromatin fiber into aggregates of loops, since only such a topology leads to such an evolutionary stable motive.

Fig. 4: *Drosophila melanogaster* (A-C) and *Arabidopsis thaliana* (D, E) multi-scaling and fine-structure (C, E), surviving averaging (B).

Fig. 5: Saccharomyces cerevisae and Schizosaccharomyces pombe multi-scaling and fine-structure.

TREE CONSTRUCTION

The fine-structured multi-scaling correlation behavi-ours show a higher similarity within different Euka-riotic species than between major regna (Fig. 6). Archaea and Bacteria sequences can be visually subdivided into different classes (Fig. 13). Pearson correlation matrices between the C(l) or δ(l) were computed, transformed into distance matrices and used for tree construction by neighbour joining (NJ) or the unweighted pair group method by average (UPGMA) using the programme PAUP. Whereas, the Eukarya tree agrees with the phylogentic tree, for Archaea and Bacteria the four qualitive groups are validated and suggest a new classification system.

Fig. 6: Comparison of the average correlation behaviour of Eukarya, Archaea and Bacteria classes.

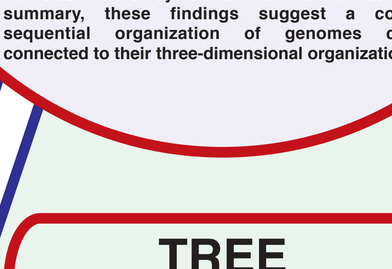
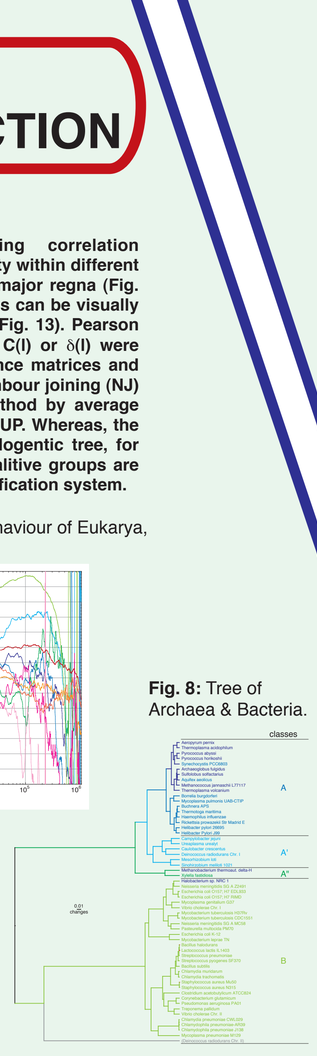


Fig. 7: Tree of Eukarya.

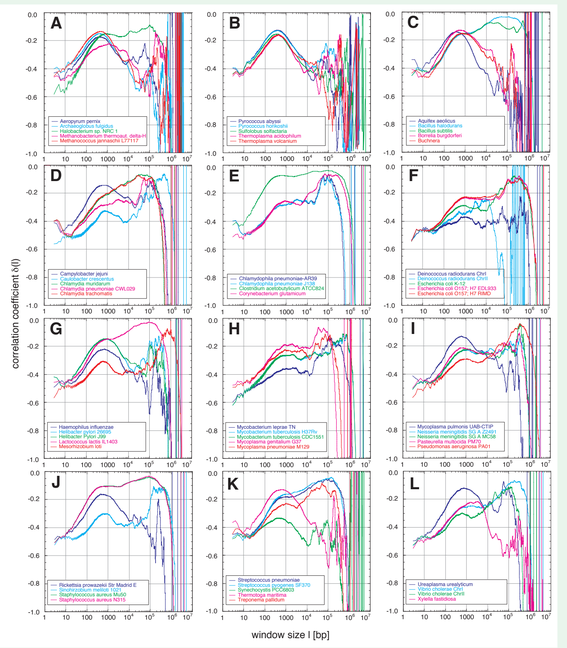
Fig. 8: Tree of Archaea & Bacteria.



ARCHAEA & BACTERIA GENOMES

All analysed Archaea (Fig. 13A-B) and Bacteria (Fig. 13C-L) genomes show beyond a simple long-range power-law, also different local correlation coefficients: A first maximum between 500 and 2000 bp is followed by a region of one or more second maxima or a plateau between ~10^5 and ~10^6 bp. The behaviour is more diverse than expected from the similarity between the chromosomes within the single Eukarya species. Nevertheless, their general appearance allows qualitative and quantitative classification by tree construction into four major classes referred to as A (e. g. *Aquifex aeolicus*, *Eropyrum pernix*), A' (e. g. *Campylobacter jejuni*), A'' (only *Methanobacterium thermoautotrophicum deltaH* and *Xylella fastidiosa*) and B (e. g. *Bacillus halodurans* or *Clostridium acetobutylicum*) with distinct general behaviour. Amazingly, in the lattest the degree of correlation sometimes is extreme and even above -0.1 and decrease for window sizes >10^5 bp sharply with hardly any fluctuation. This supports again the high correlation degree suppressing the fluctuations, common else and for random sequences (Fig. 1B).

Fig. 13: Archaea (A, B) and Bacteria (C-L) multi-scaling and fine-structure.



Literature

Avnir, D., editor, The fractal approach to heterogeneous chemistry. *John Wiley & Sons*, 1989.

Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J. F. & Arneodo, A. Long-range correlations in genomic DNA: A signature of the nucleosomal structure. Phys. Rev. Lett. 86 (11), 2471-2474, 2001.

Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. J. Mol. Biol. 316, 903-918, 2002.

Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture, and gene regulation in mammalian cells. Nat. Rev. Gen. 2(4),292, 2001.

Knoch, T. A. Three-dimensional organization of chromosome territories in simulation and experiments. *Diploma-Thesis* (German), Deutsches Krebsforschungszentrum, Heidelberg, Germany, and Fakulty for Physics und Astronomy, University of Heidelberg, Heidelberg, Germany, 1998.

Knoch, T. A., Münkel, C., & Langowski, J. Three-dimensional organization of chromosome territories and the human interphase nucleus. *High Performance Scientific Supercomputing*, Scientific Supercomputing Center (SSC) Karlsruhe, University of Karlsruhe (TH), editor Wilfried Juling, June 1999.

Knoch, T. A., Münkel, C., & Langowski, J. Three-dimensional organization of chromosome territories and the human interphase nucleus. *High Performance Computing in Science and Engineering* 1999, editors Egon Krause and Willi Jäger, High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3-540-66504-8, 229-238, 2000.

Knoch, T. A. Approaching the three-dimensional organization of the human genome. *Dissertation*, Ruperto Carola University, Heidelberg, Germany, and Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-009959 (soft cover 3rd ed.), ISBN 3-00-009960-3 (hard cover, 3rd ed.), 2002.

Münkel, C. & Langowski, J. Chromosome structure predicted by a polymer model. *Phys. Rev. E* 57 (5), 5888-5896, 1998.

Münkel, C., Eils, R., Dietzel, S., Zink, D., Mehring, C., Wedemann, G., Cremer, T. & Jörg Langowski Compartimentalization of Interphase Chromosomes Observed in Simulation and Experiment. *J. Mol. Biol.* 285, 1053-1065, 1999.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. Long-range correlations in nucleotide sequences. Nature 356, 168-170, 1992.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H.E. & Goldberger, A. L. Mosaic organization of DNA nucleotides. Phys. Rev. E 49 (2), 1685-1689, 1994.

Pienta, K. J. & Coffey, D. S. A structural analysis of the role of the nuclear matrix and DNA loops in the organization of the nucleus and the chromosome. Higher Order Structure in the Nucleus, edited by P. R. Cook and R. A. Laskey, *J. Cell Sci*., Supplement I,123-135, 1984

Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Goldberger, Z. D., Havlin, S., Mantegna, R. N., Ossadnik, S. M., Peng, C.-K. and Simons, M. Statistical mechanics in biology: how ubiquitous are long-range correlations? Physica A 205, 214-253, 1994.

Voss, R. F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. 68 (25), 3805-3808, 1992.

Yokota, H., van den Engh, G. J., Hearst, J. E., Sachs, R. K. and Trask, B. J. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol*. 130(6), 1239-1249, 1995.

Zirbel, R. M., Mathieu, U., Kurz, A., Cremer, T. & Lichter, P. Evidence for a nuclear compartment of transcription and splicing located at chromosome domain boundaries. *Chrom. Res*. 1, 92-106, 1993.

Approaching the Sequential and Three-Dimensional Organization of Archaea, Bacteria and Eukarya Genomes

Knoch, T. A., Göker, M. Lohner, R. & Langowski, J.

T*IGR's 14th International Genome Sequencing & Analysis Conference*, Boston, USA, 2nd - 5th September, 2002.

Abstract

The largely unresolved sequential organization, i.e. the relations within DNA sequences, and its connection to the three-dimensional organization of genomes was investigated by correlation analyses of completely sequenced chromosomes from Viroids, Archaea, Bacteria, Arabidopsis thaliana, Saccharomyces cerevisae, Schizosaccharomyces pombe, Encephalitozoon cunniculi, Drosophila melangoster, Homo sapiens, chloroplasts and mitochondria. All sequences revealed long-range power-law correlations almost on the entire observable scale. The local correlation coefficient shows close to random correlations on the scale of a few base pairs, a first maximum from 40-3400 bp, and often a region of one or more second maxima from 10^5-3x10^5 bp. This multi-scaling behaviour is species specific and can be explained by a block organization of genomes. Within this multi-scaling behaviour an additional fine-structure is present and attributable to the codon usage in all except the human sequences. Here it is connected to nucleosomal binding. Computer generated random sequences assuming a block organization, the codon usage and nucleosomal binding agree with these results. Mutation by simulated sequence reshuffling destroyed all correlations, thus their stability seems evolutionary tightly controlled and connected to the spatial genome organization. On large scales the sequence correlations agree very well with the three-dimensional folding of the 30 nm chromatin fibre into the Multi-Loop-Subcompartment (MLS) model, in which ~100 kbp loops form rosettes, connected by a linker, within chromosomes.

Corresponding author email contact: TA.Knoch@taknoch.org

Keywords:

Genome, genomics, genome organization, genome architecture, structural sequencing, architectural sequencing, systems genomics, coevolution, holistic genetics, genome mechanics, genome function, genetics, gene regulation, replication, transcription, repair, homologous recombination, simultaneous co-transfection, cell division, mitosis, metaphase, interphase, cell nucleus, nuclear structure, nuclear organization, chromatin density distribution, nuclear morphology, chromosome territories, subchromosomal domains, chromatin loop aggregates, chromatin rosettes, chromatin loops, chromatin fibre, chromatin density, persistence length, spatial distance measurement, histones, H1.0, H2A, H2B, H3, H4, mH2A1.2, DNA sequence, complete sequenced genomes, molecular transport, obstructed diffusion, anomalous diffusion, percolation, long-range correlations, fractal analysis, scaling analysis, exact yard-stick dimension, box-counting dimension, lacunarity dimension, local nuclear dimension, nuclear diffuseness, parallel super computing, grid computing, volunteer computing, Brownian Dynamics, Monte Carlo, fluorescence in situ hybridization, confocal laser scanning microscopy, fluorescence correlation spectroscopy, super resolution microscopy, spatial precision distance microscopy, auto-fluorescent proteins, CFP, GFP, YFP, DsRed, fusion protein, in vivo labelling.

Literature References

Knoch, T. A. Dreidimensionale Organisation von Chromosomen-Domänen in Simulation und Experiment. (Three-dimensional organization of chromosome domains in simulation and experiment.) *Diploma Thesis*, Faculty for Physics and Astronomy, Ruperto-Carola University, Heidelberg, Germany, 1998, and TAK Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-010685-5 and ISBN 978-3-00-010685-9 (soft cover, 2rd ed.), ISBN 3-00-035857-9 and ISBN 978-3-00-0358857-0 (hard cover, 2rd ed.), ISBN 3-00-035858-7, and ISBN 978-3-00-035858-6 (DVD, 2rd ed.), 1998.

Knoch, T. A., Münkel, C. & Langowski, J. Three-dimensional organization of chromosome territories and the human cell nucleus - about the structure of a self replicating nano fabrication site. *Foresight Institute - Article Archive*, Foresight Institute, Palo Alto, *CA, USA,* http://www. foresight.org, 1- 6, 1998.

Knoch, T. A., Münkel, C. & Langowski, J. Three-Dimensional Organization of Chromosome Territories and the Human Interphase Nucleus. *High Performance Scientific Supercomputing*, editor Wilfried Juling, Scientific Supercomputing Center (SSC) Karlsruhe, University of Karlsruhe (TH), 27- 29, 1999.

Knoch, T. A., Münkel, C. & Langowski, J. Three-dimensional organization of chromosome territories in the human interphase nucleus. *High Performance Computing in Science and Engineering 1999*, editors Krause, E. & Jäger, W., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3-540-66504-8, 229-238, 2000.

Bestvater, F., **Knoch, T. A.**, Langowski, J. & Spiess, E. GFP-Walking: Artificial construct conversions caused by simultaneous cotransfection. *BioTechniques* 32(4), 844-854, 2002.

Knoch, T. A. (editor), Backes, M., Baumgärtner, V., Eysel, G., Fehrenbach, H., Göker, M., Hampl, J., Hampl, U., Hartmann, D., Hitzelberger, H., Nambena, J., Rehberg, U., Schmidt, S., Weber, A., & Weidemann, T. Humanökologische Perspectiven Wechsel - Festschrift zu Ehren des 70. Geburtstags von Prof. Dr. Kurt Egger. Human Ecology Working Group, Ruperto-Carola University of Heidelberg, Heidelberg, Germany, 2002.

Knoch, T. A. Approaching the three-dimensional organization of the human genome: structural-, scaling- and dynamic properties in the simulation of interphase chromosomes and cell nuclei, long- range correlations in complete genomes, *in vivo* quantification of the chromatin distribution, construct conversions in simultaneous co-transfections. *Dissertation*, Ruperto-Carola University, Heidelberg, Germany, and TAK†Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-009959-X and ISBN 978-3-00-009959-5 (soft cover, 3rd ed.), ISBN 3-00-009960-3 and ISBN 978-3-00-009960-1 (hard cover, 3rd ed.), ISBN 3-00-035856-9 and ISBN 978-3-00-010685-9 (DVD, 3rd ed.) 2002.

Knoch, T. A. Towards a holistic understanding of the human genome by determination and integration of its sequential and three-dimensional organization. *High Performance Computing in Science and Engineering 2003*, editors Krause, E., Jäger, W. & Resch, M., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3- 540-40850-9, 421-440, 2003.

Wachsmuth, M., Weidemann, T., Müller, G., Urs W. Hoffmann-Rohrer, **Knoch, T. A.**, Waldeck, W. & Langowski, J. Analyzing intracellular binding and diffusion with continuous fluorescence photobleaching. *Biophys. J.* 84(5), 3353-3363, 2003.

Weidemann, T., Wachsmuth, M., **Knoch, T. A.**, Müller, G., Waldeck, W. & Langowski, J. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. *J. Mol. Biol.* 334(2), 229-240, 2003.