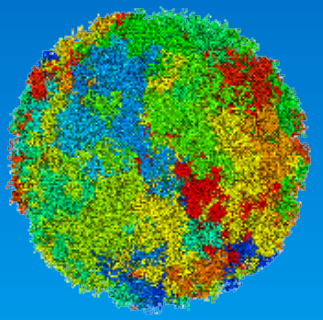


From Sequence to Morphology

-

Towards a Holistic Understanding

of the

Organization of the Human Genome

-

Long-Range Correlations

in

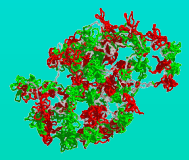
Complete Sequenced Genomes

by

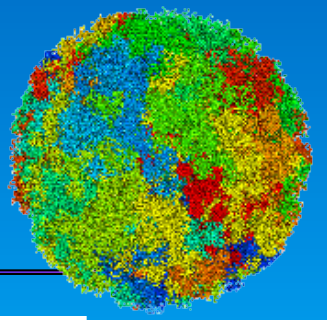
Tobias A. Knoch

Kirchhoff Institut for Physics, Ruperto-Carola University
Deutsches Krebsforschungszentrum (DKFZ)
Heidelberg, Germany

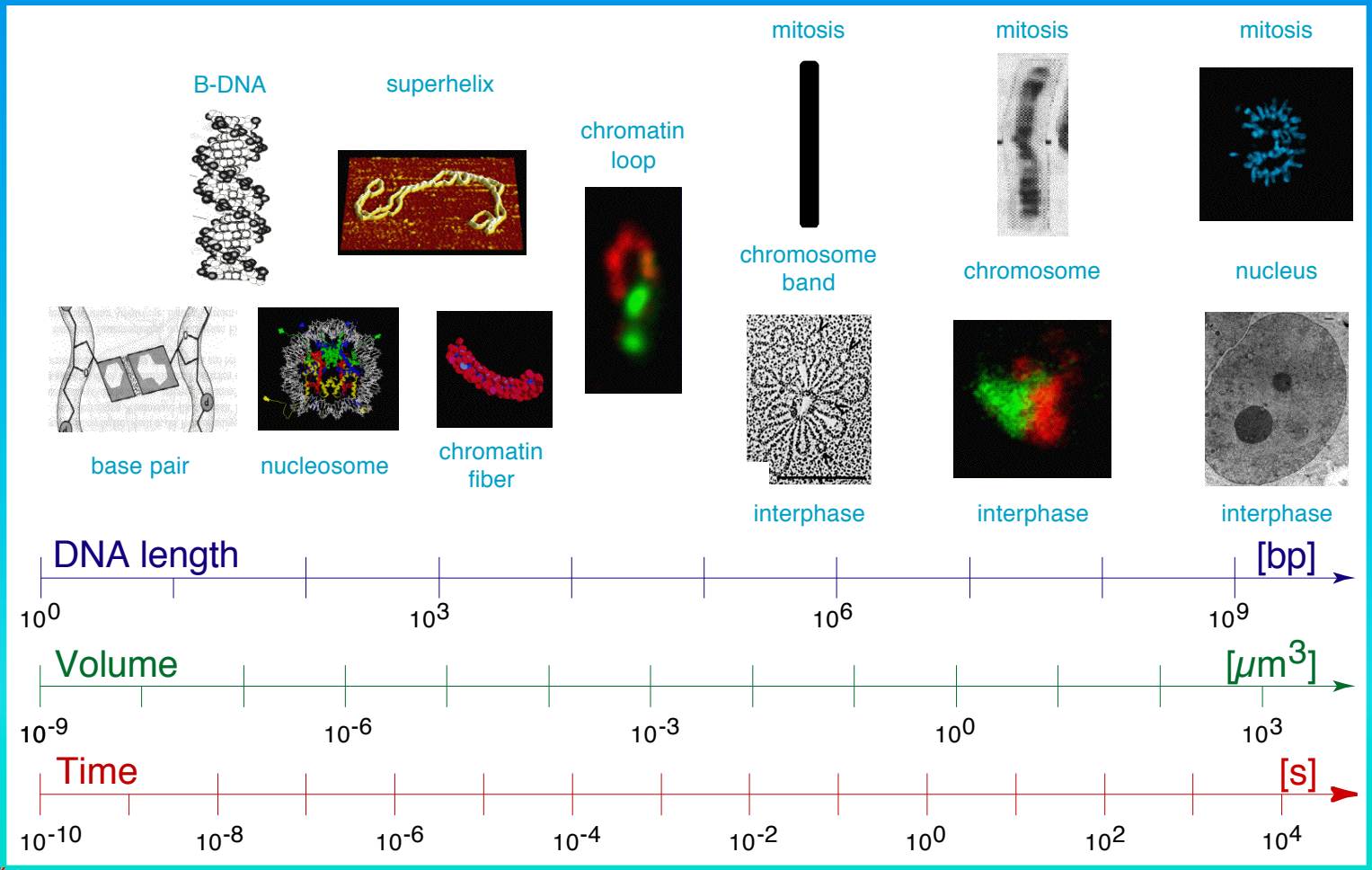
Erasmus MC, Rotterdam, The Netherlands

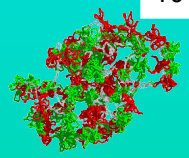


Dynamic and Hierarchical Genome Organization

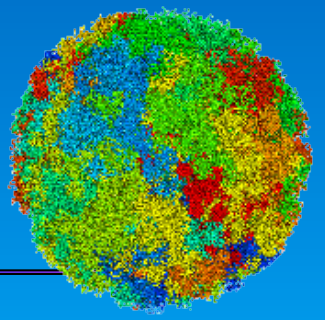


10 and 13 orders of magnitude concerning length and time scales are bridged.
Are and how are all of these organization levels connected to fullfill their obvious functions, e. g. gene regulation or replication, since they are optimized by evolution ?





Sequential Organization of Genomes



Determination of the concentration fluctuation function $C(l)$ and its local slope the correlation coefficient $\delta(l)$ reveal multi-scaling long-range correlation up to 10^6 to 10^7 bp in *Homo sapiens* which clearly deviate from random sequences with high significance (decreasing the nearer to the cut-off).

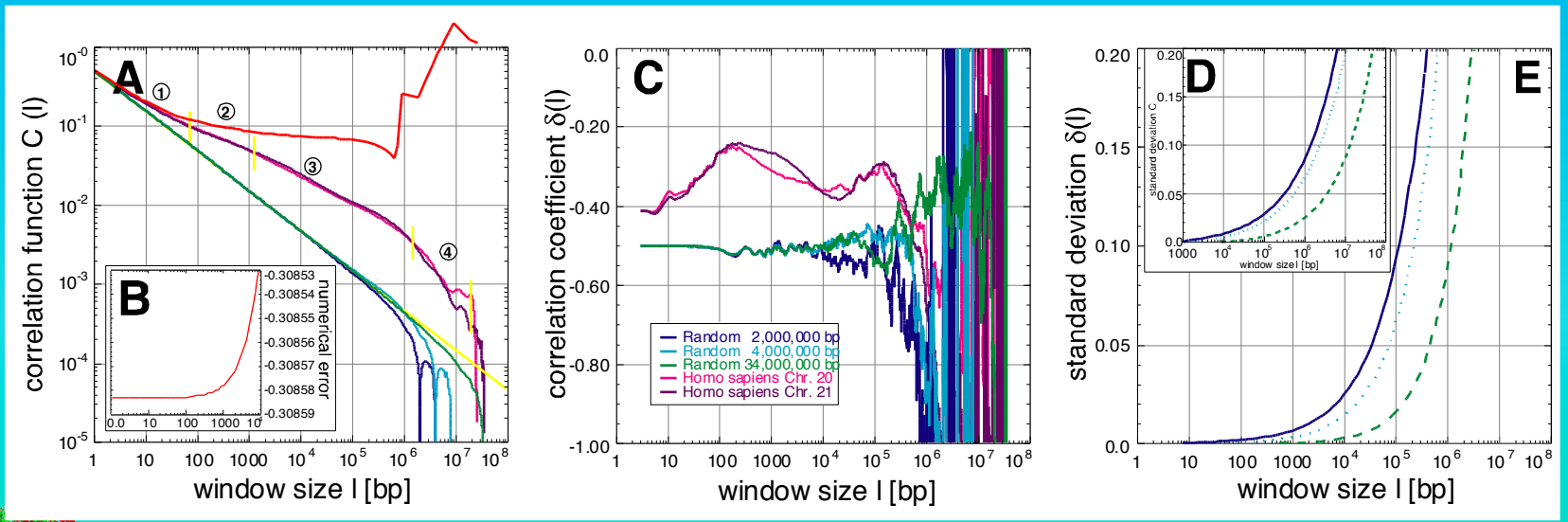
On large scales this might only be due to a strong and definite three-dimensional genome organization.

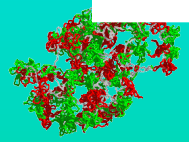
$$C(l) = \sqrt{\left\langle (c_l - \bar{c}_L)^2 \right\rangle_s}$$

numerically unstable

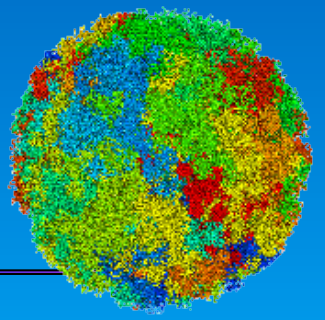
$$C(l) = \sqrt{\frac{1}{L-l+1} \sum_{s=1}^{L-l} \left(\frac{1}{l} \sum_{k=1}^{l} n - \frac{1}{L} \sum_{k=1}^{L} N \right)^2}$$

numerically stable





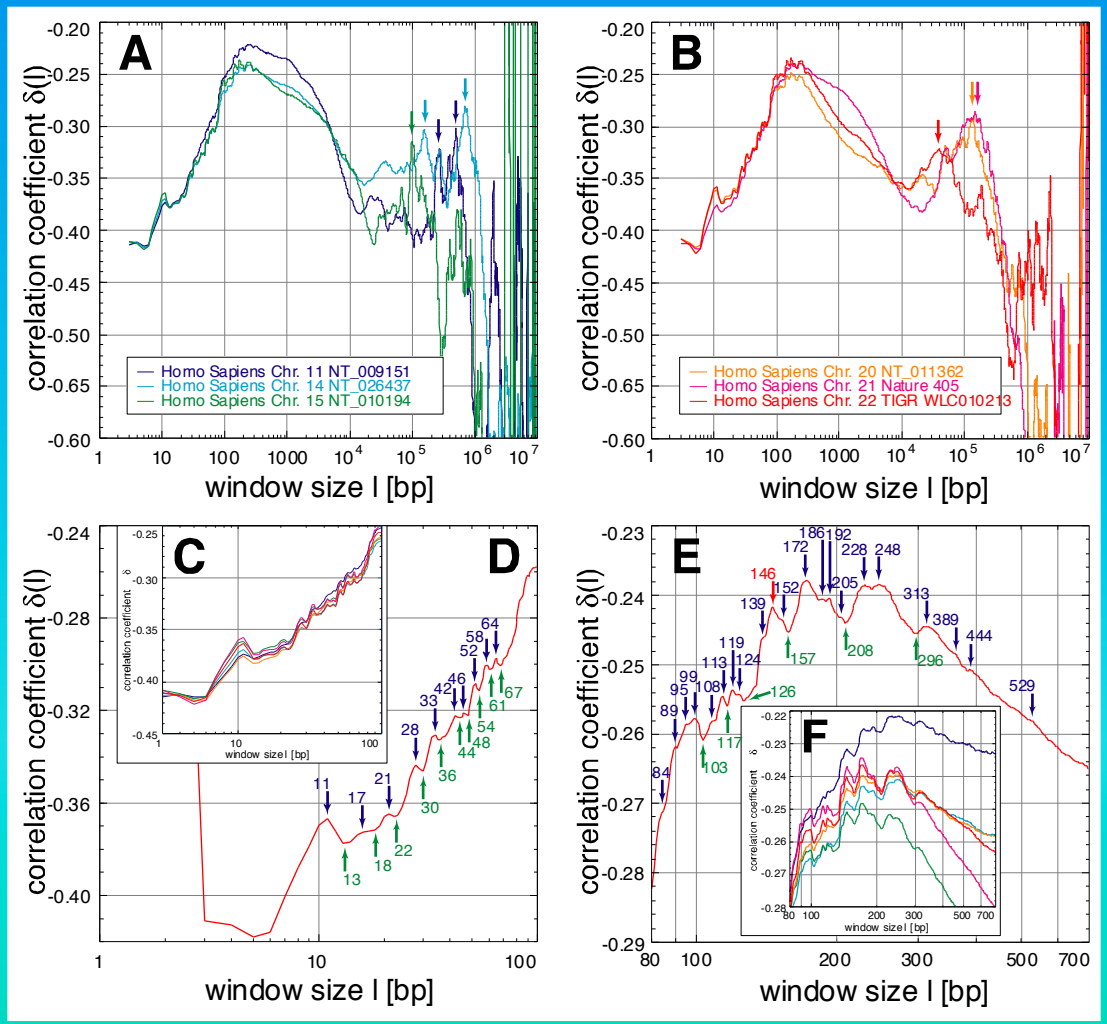
Fine-Structured Multi-Scaling Long-Range Correlations of *Homo sapiens*



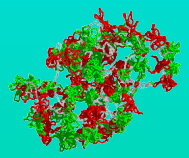
The general behaveour is characterized by first maximum of the correlation coefficient $\delta(l)$ at ~250 bp and at $1x10^5$ to $3x10^5$ bp, both due to a globular block structure of genomes. Due to their fine structure the first is attributable to nucleosomal binding and the latter due to aggregation of chromatin loops as in the MLS model.

general behaviour

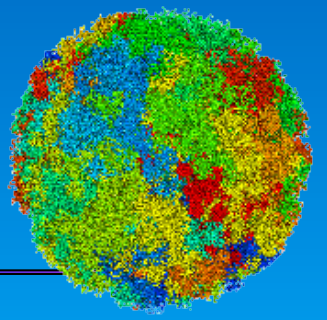
fine structure



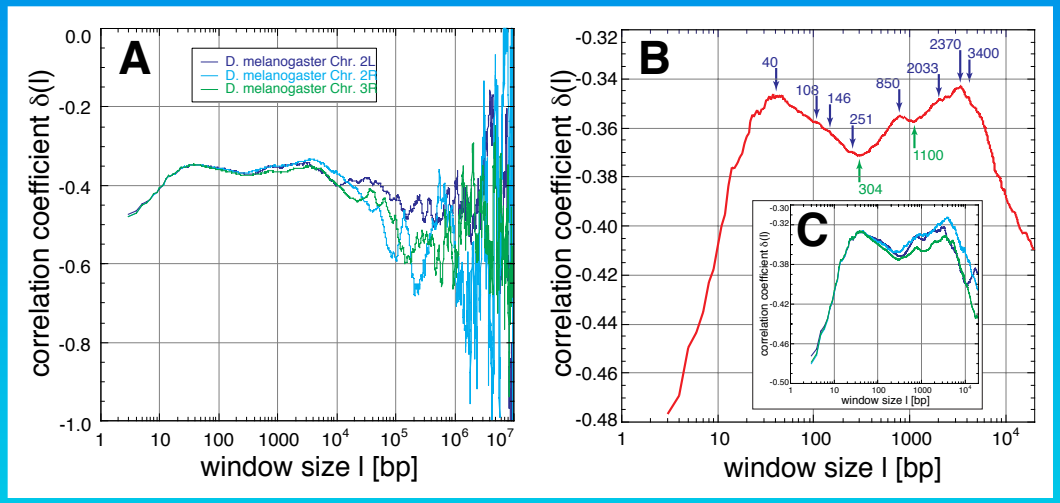
the fine structure survives averaging over several human chromosomes.



Fine-Structured Multi-Scaling Long-Range Correlations of *Drosophila melanogaster* and *Arabidopsis thaliana*

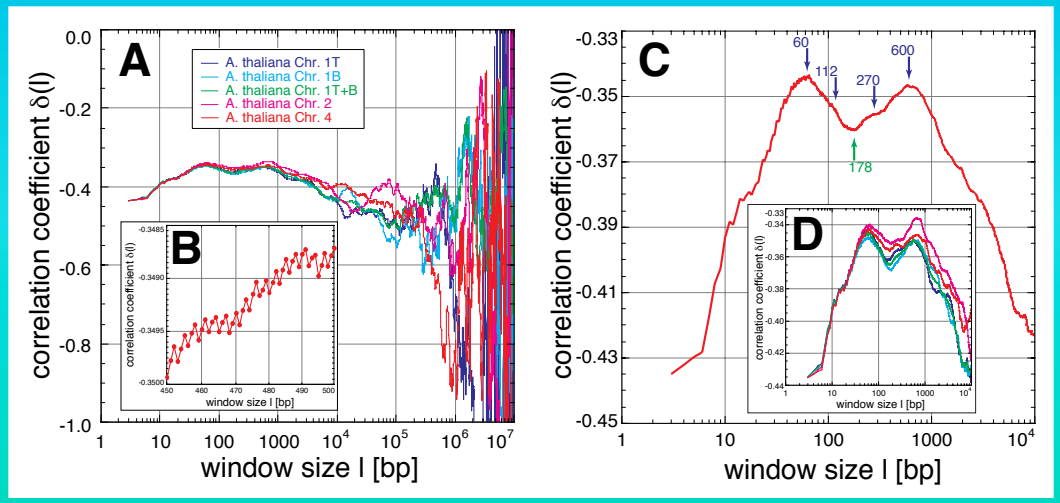


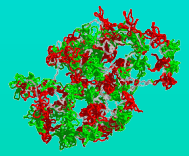
The general behaveour is characterized by two main submaxima between ~40 and ~3000 bp, again due to a block structure. Two fine structures are present: the main strong periodicity of 3 is attributable to the codon usage and a minor one is attributabe to nucleosome binding. Ultra-structural fine-structures seem plausible.



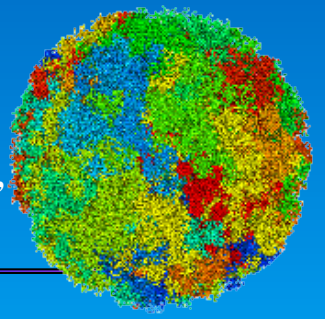
Drosophila melanogaster

Arabidopsis thaliana

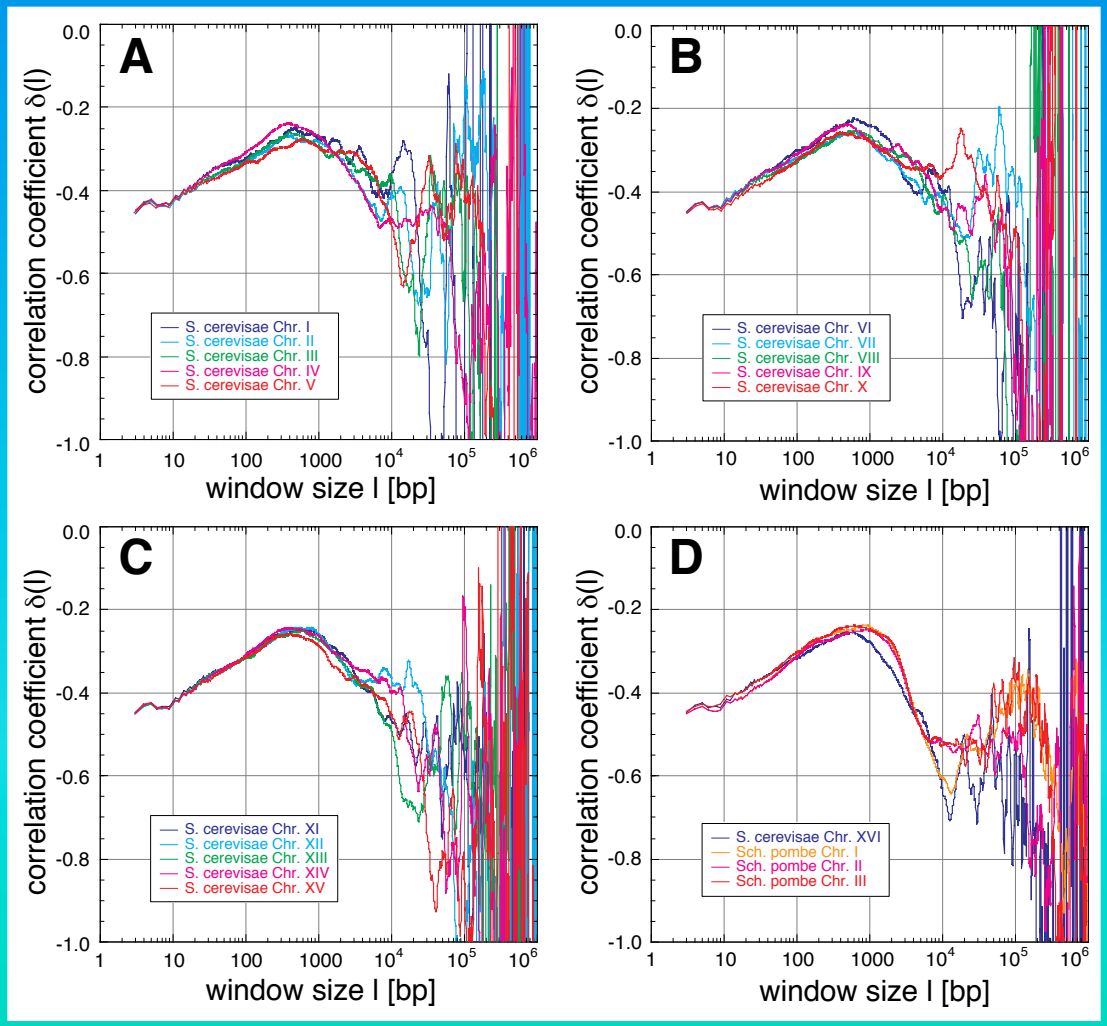


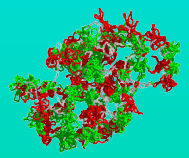


Fine-Structured Multi-Scaling Long-Range Correlations of *Saccharomyces cerevisae and Schizosaccharomyces pombe*

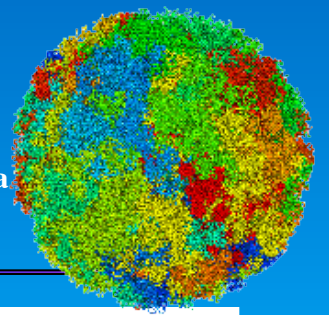


The general behaveour is characterized by a maximum of the correlation coefficient $\delta(l)$ at ~500 bp and ~900 bp, respectively, both due to a globular block structure of genomes. Their fine structure is attributable to the codon usage. Utra-structural fine-structures might seem plausible.

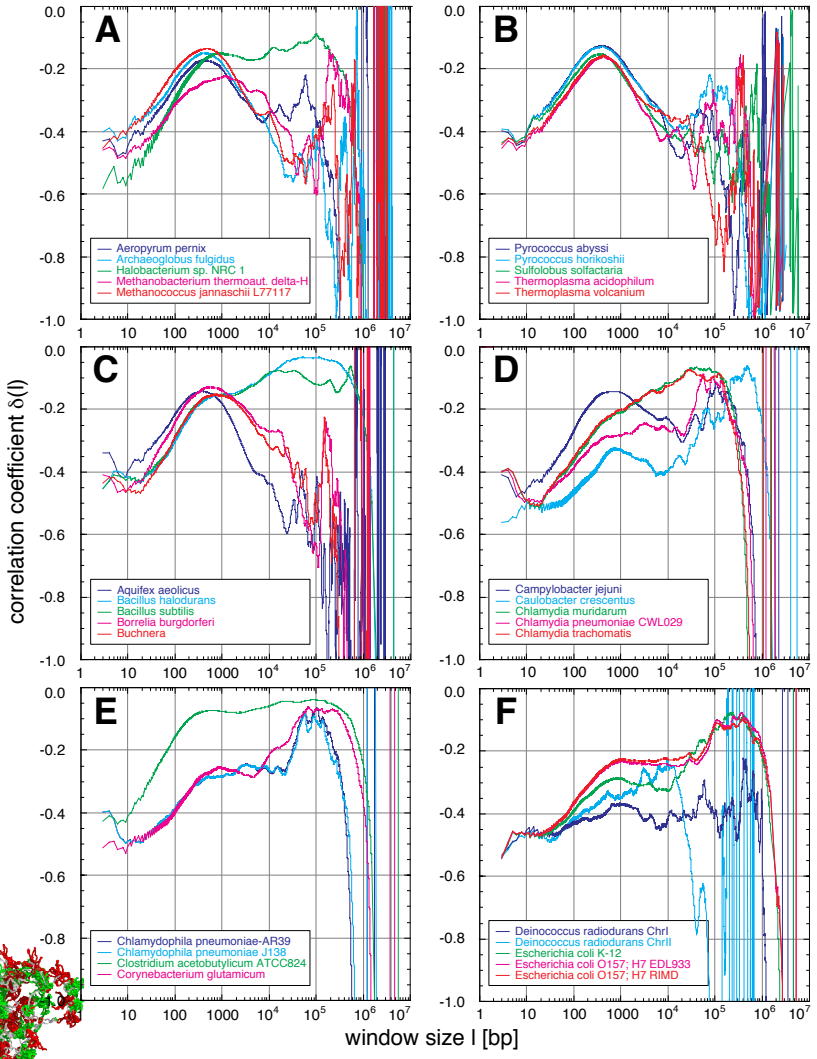


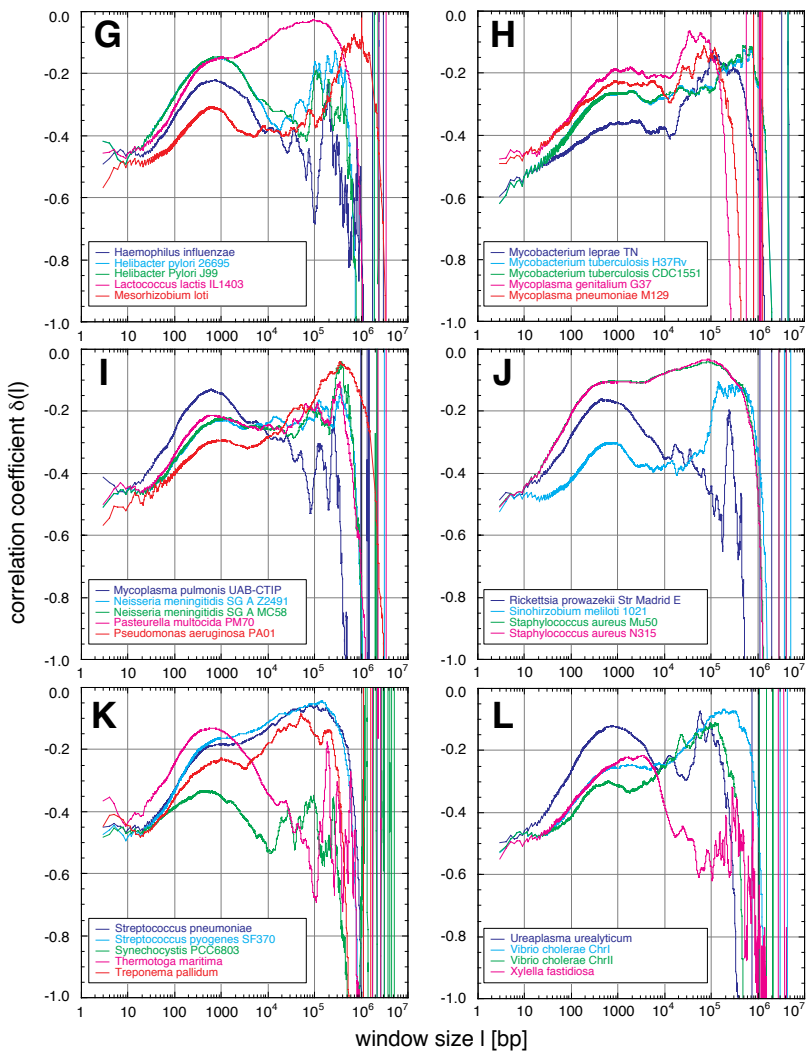


Fine-Structured Multi-Scaling Long-Range Correlations of *Archaea and Bacteria*

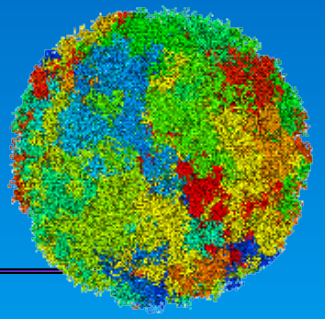


The general behaveour reveals four major classes characterized by a first maximum ~ 1x10^3 bp and sometimes a second maximum both associated with a block organization of genomes. Archaea and extremophiles, however, have mainly only the first maximum. The fine structure is codon usage and ultra-structural associated.



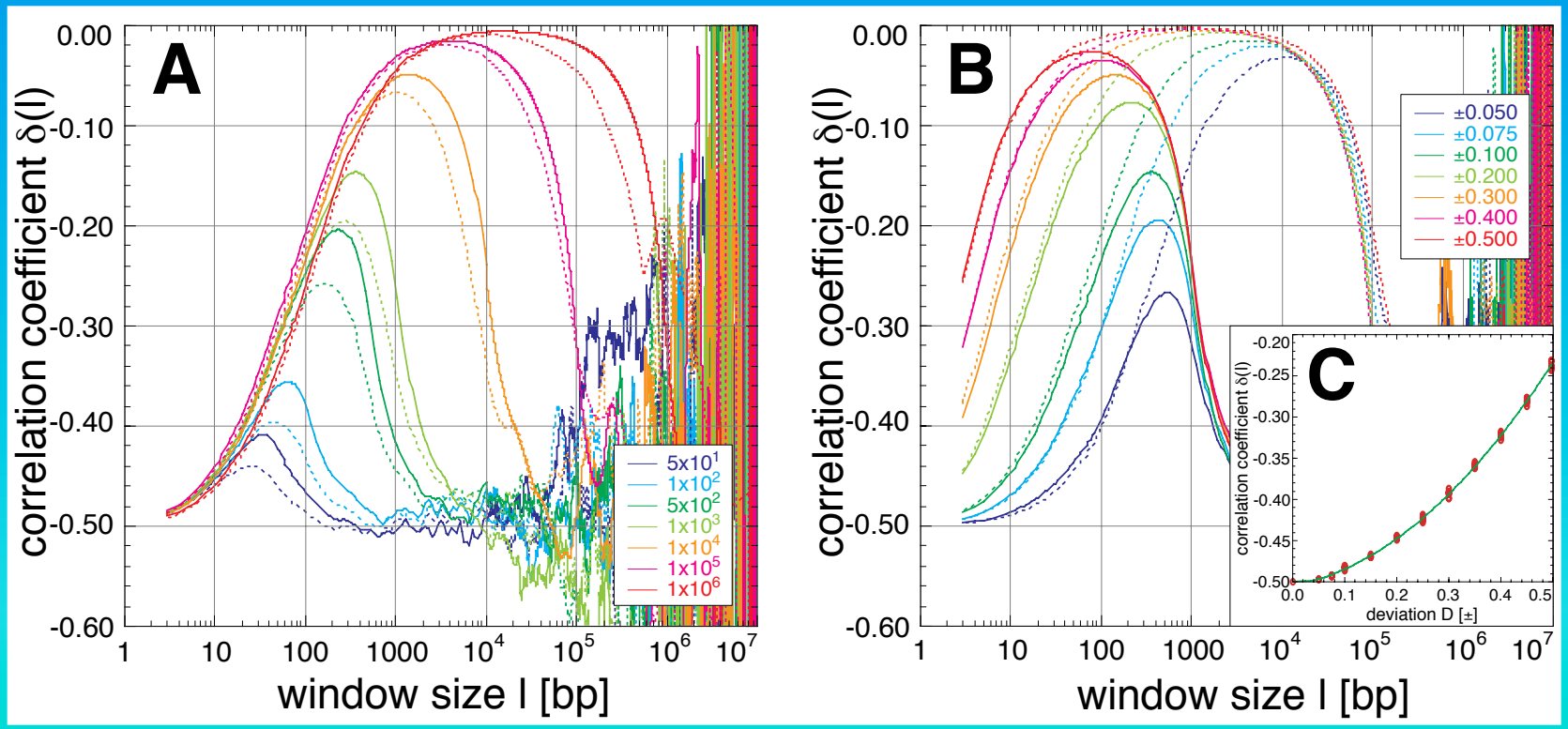


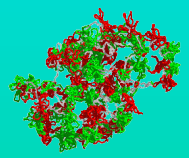
Simulation of the Block Structure of Genomes



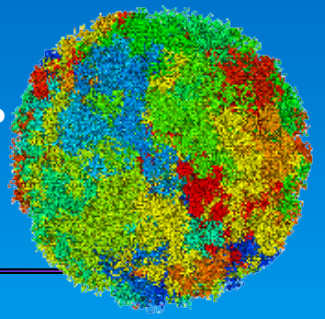
Arteficial sequences from blocks of random length B from $[B\pm10\%]$ or $[0, B]$ and different nucleotide composition lead to a global maximum whose position and height are proportional to the block length (A). The nucleotide concentration D has an inverse effect (B) and follows a quadratic behaveour for $\delta(l{=}3,D) = -0.5 + 0.113D + 0.855D^2$(C). A fine-structure is not present either on a local or a global scale - contrasting the with extreem correlation degree and smoothness should be explicitely stressed.

Thus, the general coorelation behaveour of genomes can be explained by a block organization of genomes.



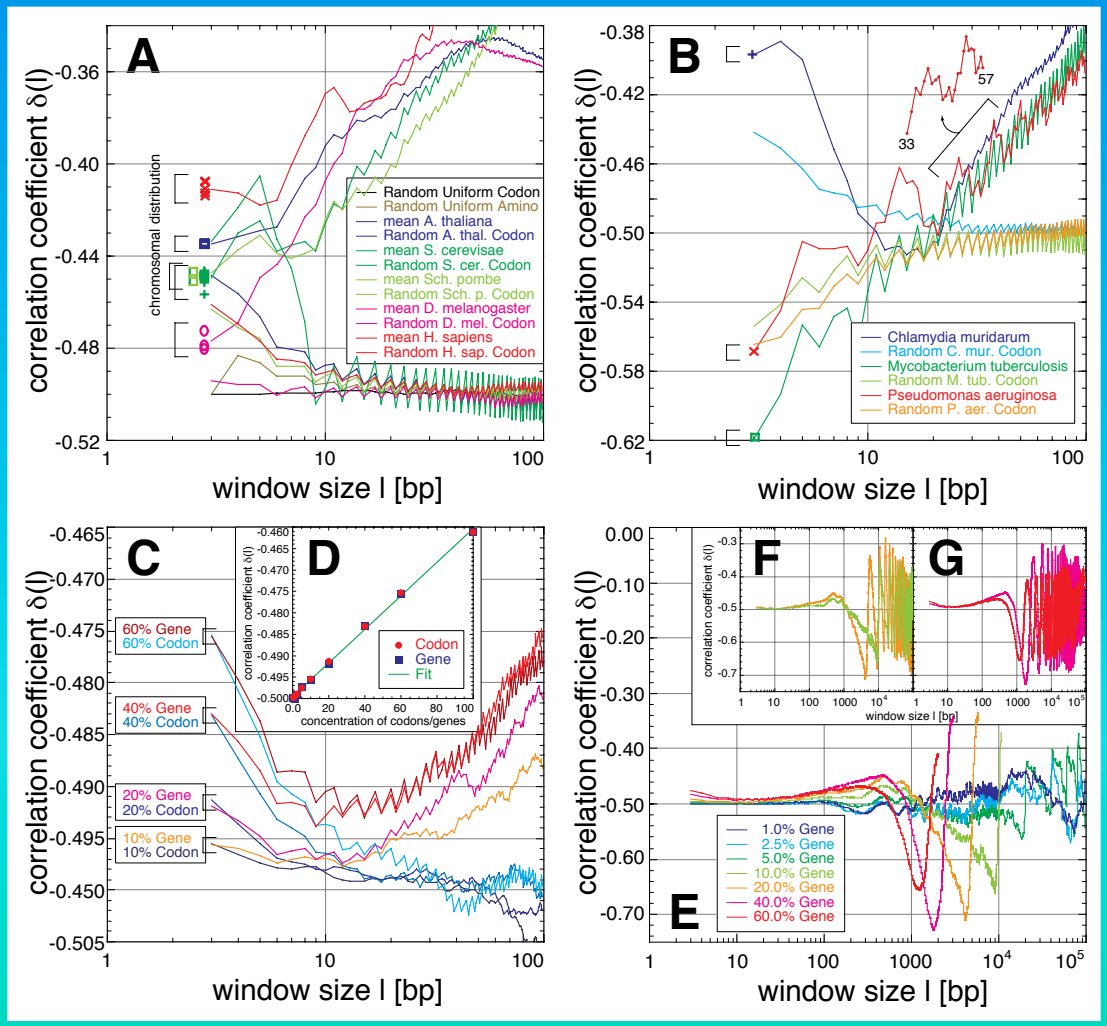


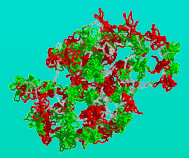
Simulation of the Codon and Gene Fine Structure of Genomes



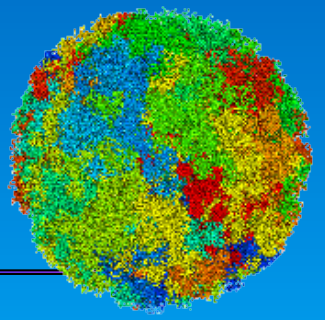
Arteficial sequences with a uniform distribution of the 20 amino acids already leads to a periodicity of 3 bp up to large l (A,B). Its appearance, sarting behaveour at $l = 3$ bp and persistence is related to the specie specific codon usage (A,B) and the concentration c of codons ($\delta(l=3,c) = -0.5 + 0.046c$) within an arteficial random sequence and are stronger if codons are arranged in a gene/block like fashion (C, D). Organization of codons in genes/ blocks leads to a maximum of $\delta(l)$ and oscillations due to the gene/block length.

Thus, the fine structures of genomes are partly due to the codon ussage and gene organization of genomes.



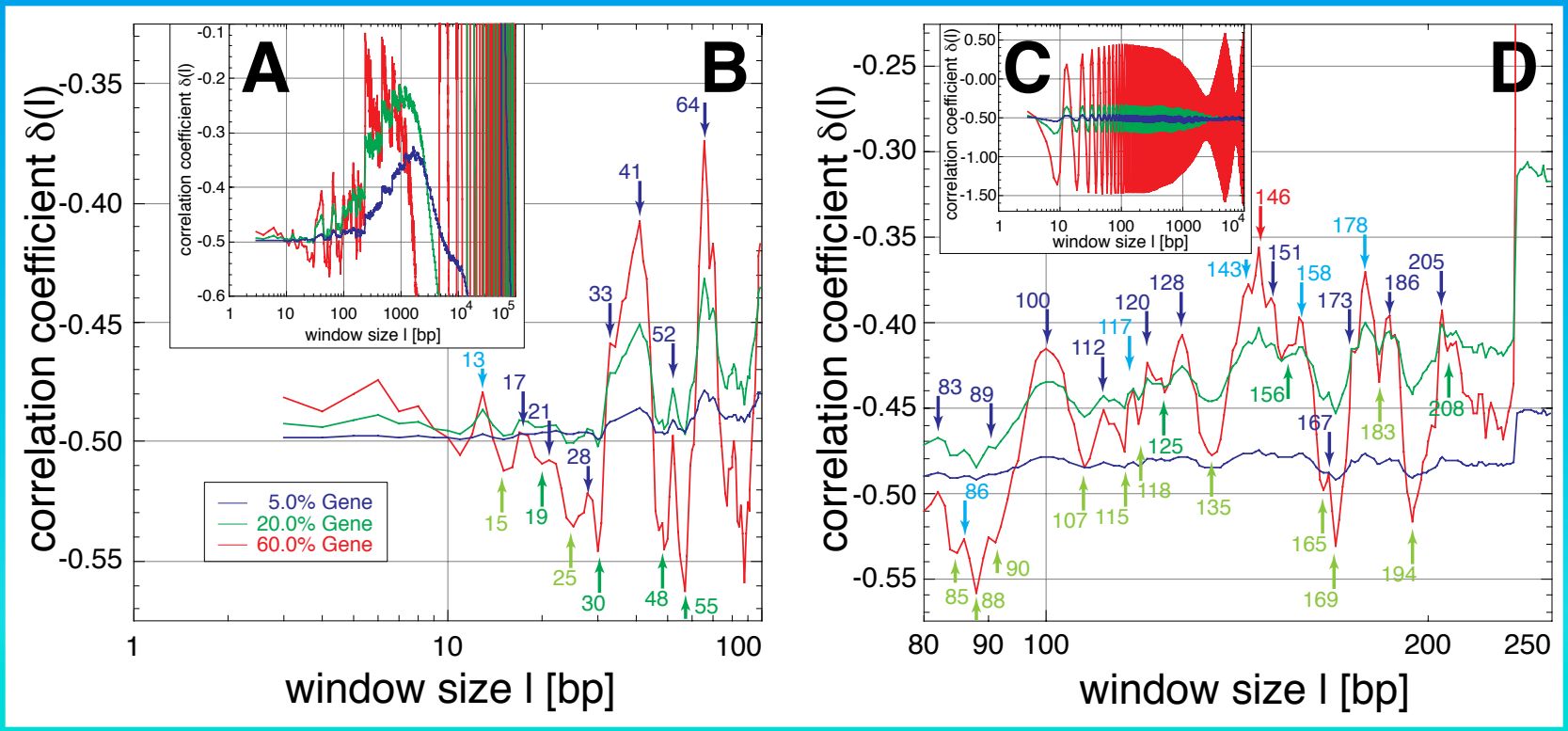


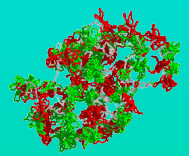
Simulation of the Nucleosomal Fine Structure of Genomes



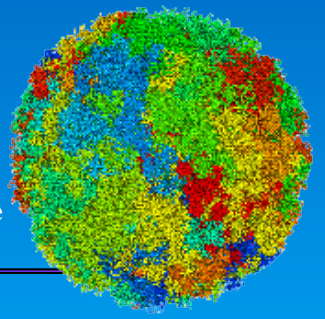
Arteficial sequences using nucleosome binding sequences as building blocks and organizing them in repeats within blocks/genes results in large agreement of local maxima and minima found in the sequences of *Homo sapiens* and depend on the concentration of blocks/genes in the arteficial sequence (A, B, D).Use of a mixture between two different binding sequences results in highly ordered periodicities of 10 bp attributed to the helical pitch (C). The general behaveour is according to the block/gene like organization.

Thus, the fine structures of genomes are partly due to the nucleosomes and periodicities within genomes.



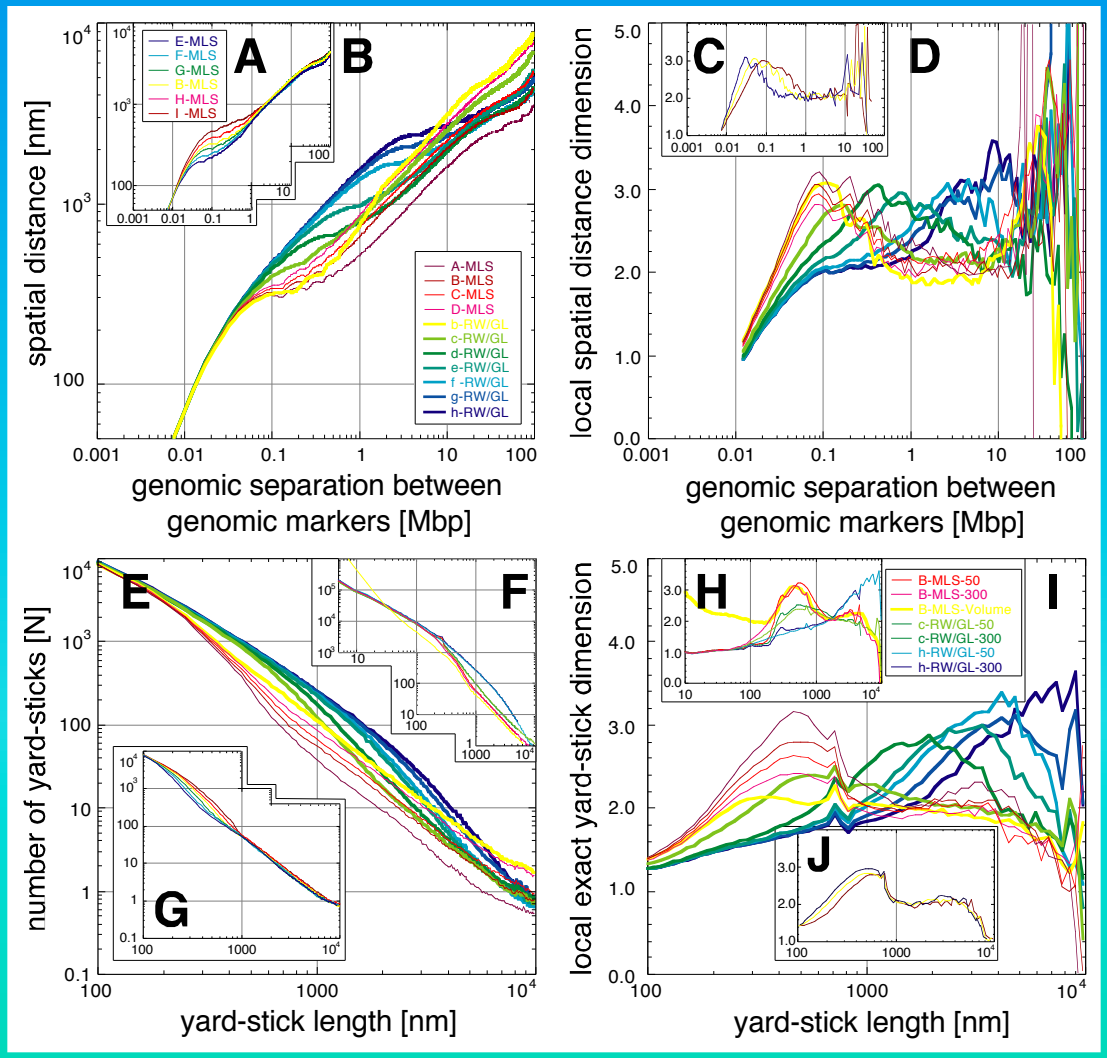


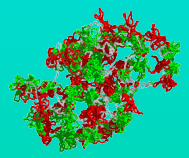
Scaling of Structural Three-Dimensiona Genom Organization Reveals the Large-Scale Fine Structure of *Homo sapiens* Genomes



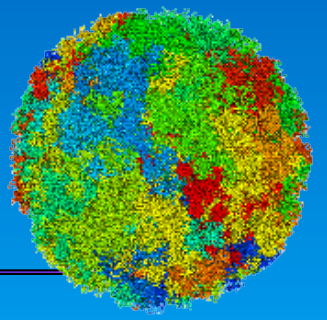
Scaling analysis of the Random-Walk/Giant-Loop (RW/GL) and the Multi-Loop Subcompartment (MLS) models of higher chromatin organization reveal a distinct model dependent scaling behaveour similar to the fine structures of the correlation behaveour found for *Homo sapiens* on scales of $5x10^4$ to $3x10^5$ bp.

Thus, keeping the notion that - what is near in sequence space should also be near in 3D space - in mind, the sequential is closely related to the 3D organization of genomes.

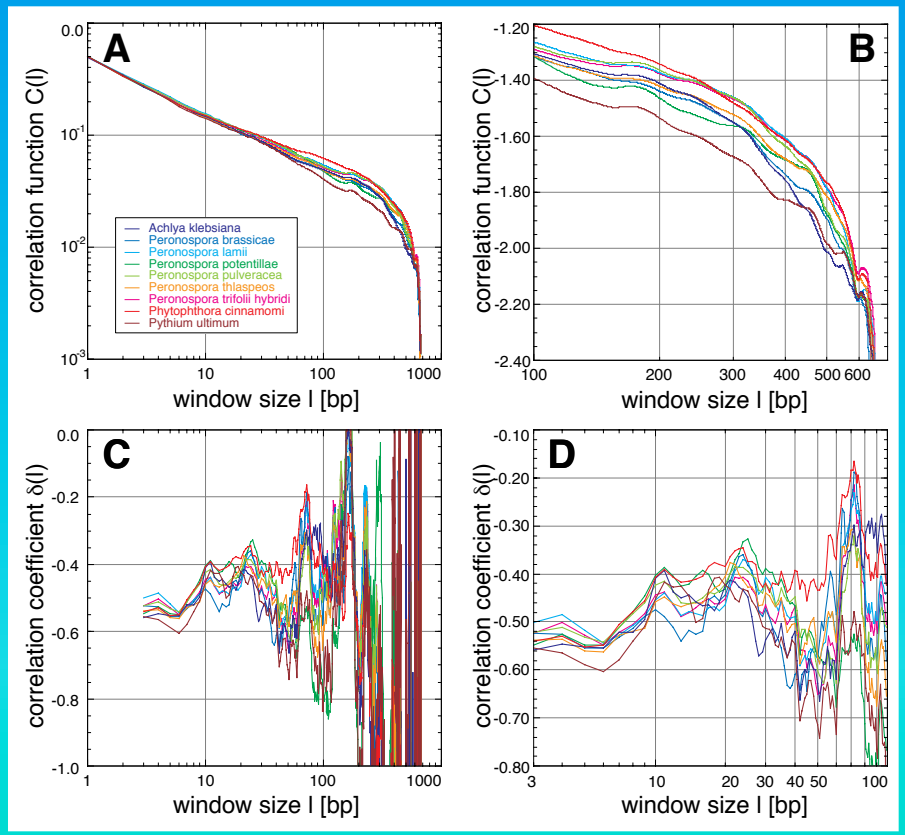


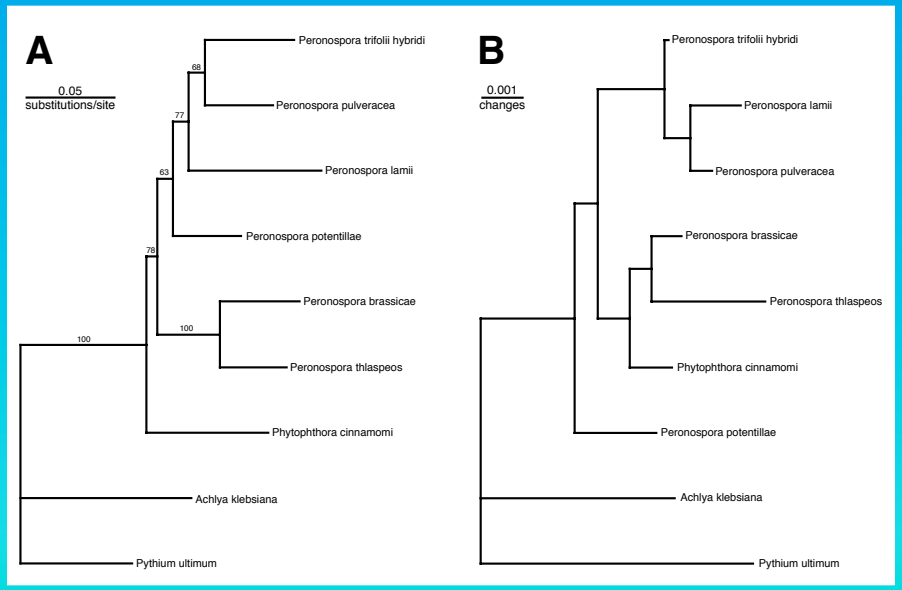


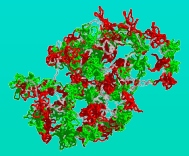
Correlations in β-Tubulin Genes of Oomycetes, Tree-Construction and Comparison to Phylogenetic Trees



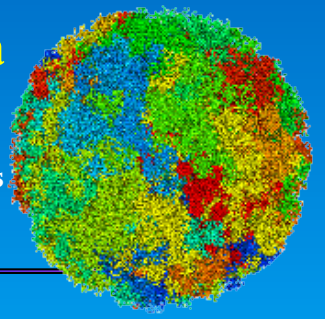
Although correlation analysis seems to reveal mostly random behaveour a more detailed inspection reveals clear relations between the sequences which are due to the similarities of sequence (A-D, left) and leads to the known phylogenetic relations (B, right) which also can be quantitatively assessed (A, right).





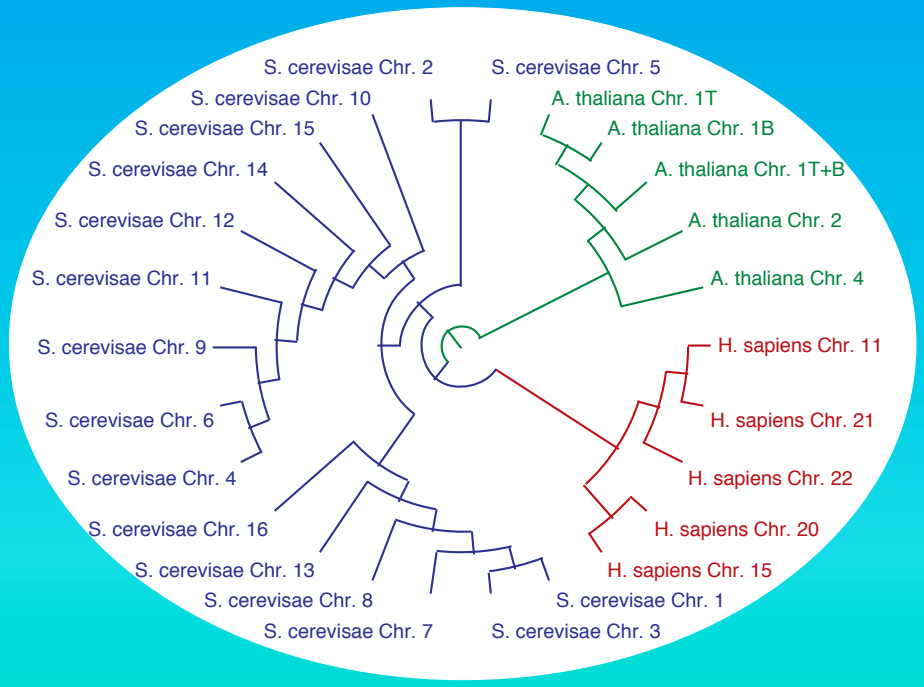


Tree Construction using the Correlation Coefficient $\delta(l)$ of Eukarya as well as Archaea and Bacteria

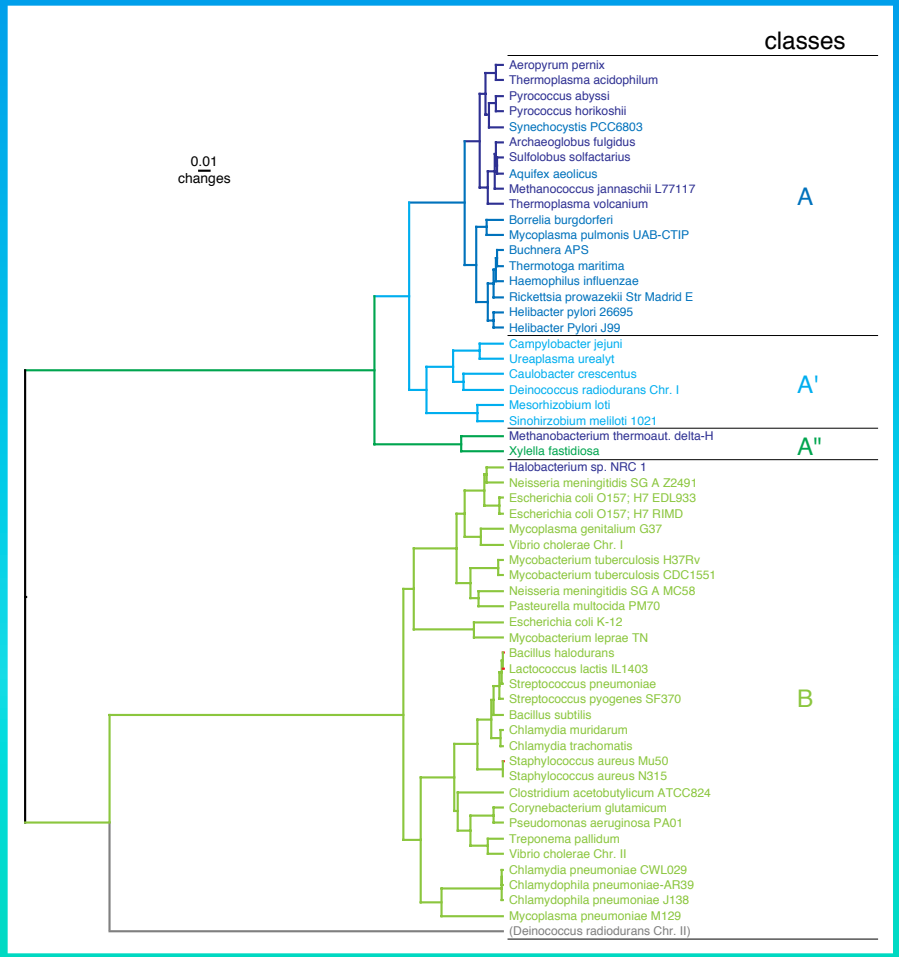


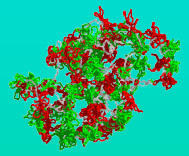
Tree construction from Eukarya reveals proper specie separation although the chromosome order within species is neither due to chromosome length, nucleotide composition and remains unclear. Tree construction of Archea and Bacteria reveals proper separation into four morphologically distinct groups.

Eukarya

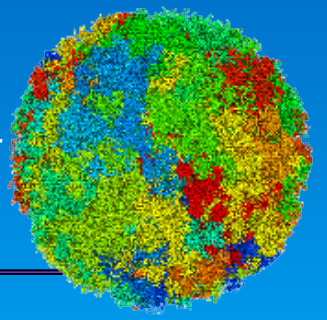


Archaea and Bacteria



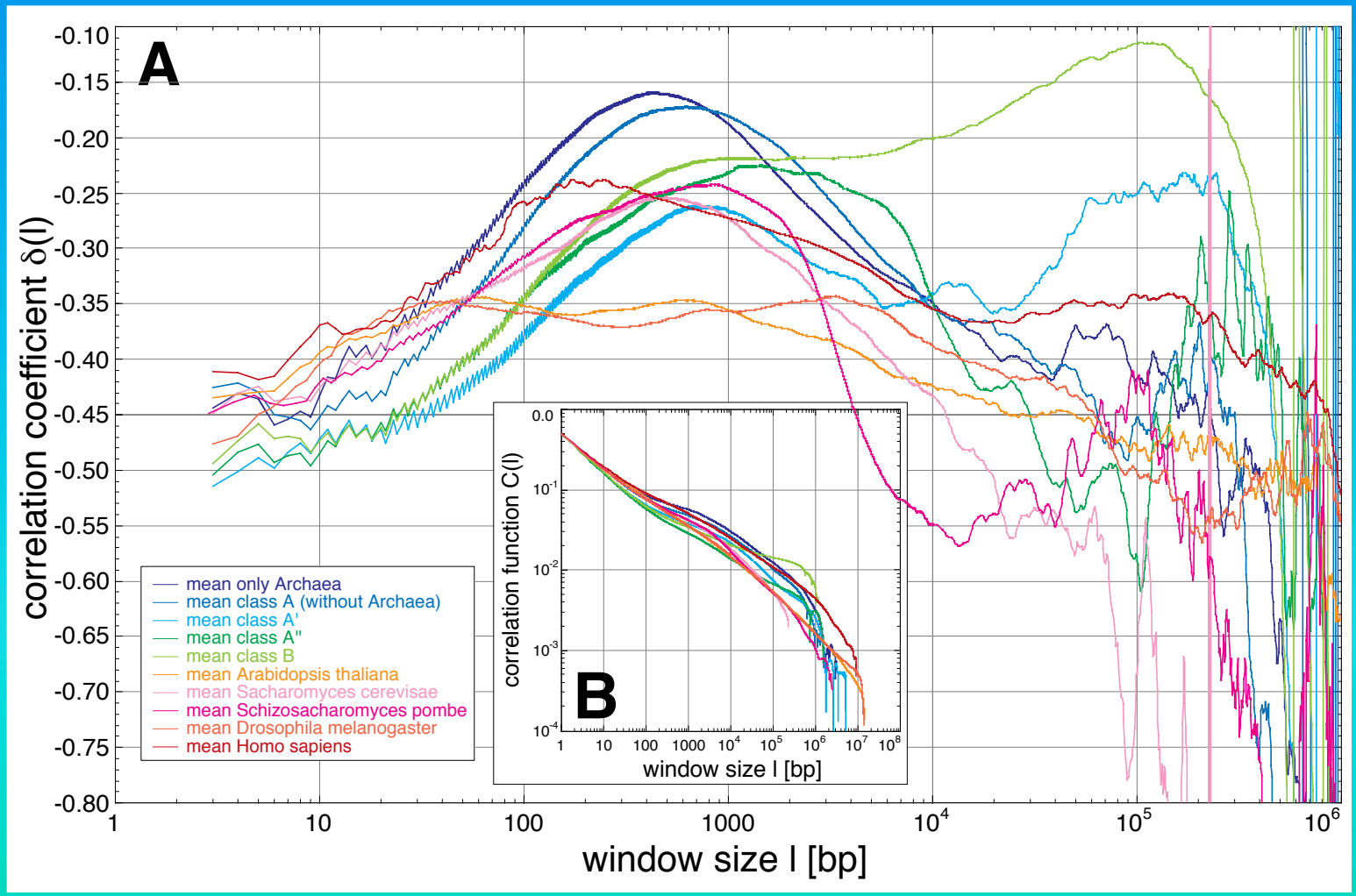


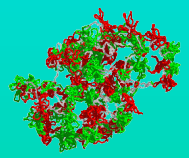
Correlation Comparison of Averages of Analysed Genomes



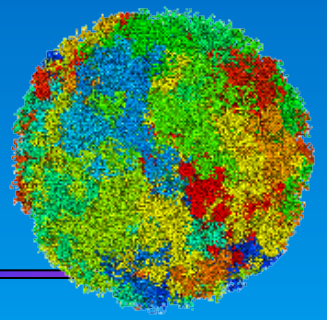
Averages over the different chromosomes within Eukarya and over the four different morphologically and tree-constructively Archaea and Bacteria trees reveals on the one hand the common behaveour within one specie/ group and on the other hand the great differences between species/groups. The different fine structures are also clearly visible for the different Eukarya, Archaea and Bacteria.

Thus, the sequential organization is far more complex than previously thought.





Conclusion



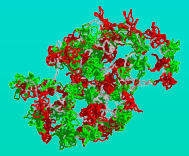
The sequential and three-dimensional organization of genomes is tightly interconnecte and thus genomes can only be understood in their complexity in a holistic manner.

- **Long-range power-law correlations were found on almost the entire observable scale of completely sequenced genomes, i.e. 0.5×10^6 to 3.0×10^7 bp.**

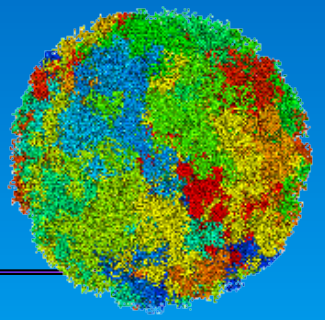
- **The general behaveour of these correlations is characterized by specie specific multi-scaling starting showing maxima at ~40 bp to 3400 bp and 1.0×10^5 bp to 3.0×10^5 which can be explained by a block organization of genomes.**

- **Within the multi-scaling behaveour three-distinct fine structures appear:**
  - **a) A periodicity of three from local to mid-ranged scales attributable to the codon usage (a sequential cause).**
  - **b) A fine-structure on mid-ranged scales attributable to the nucleosome and its binding (a mixture between a sequential and structural cause).**
  - **c) A fine structure on large scales most likely due to a rosette like 3D chromatin organization and associated periodicities thereof (mainly a structural cause).**

- **Correlation trees can be constructed from correlation coefficients and can be compared to e.g. phylogenetic trees. Species can be separated and morphologic as well as quantitative analysis leads to a new classification system for Archaea and Bacteria.**



Acknowledgements



Biophysics of Macromolecules DKFZ

Thomas Weidemann

Gabriele Müller

Waldemar Waldeck

Jörg Langowski

Molecular Biophysics KIP

Katalin Fejes-Toth

Malte Wachsmuth

Karsten Rippe

Biomedical Structure Analysis DKFZ

Felix Bestvater

Eberhard Spiess

The Cremer Labs

Joachim Rauch

Irina Solovei

Michael Hausmann

Christoph Cremer

Thomas Cremer

Molecular Genetics DKFZ

Karsten Richter

Peter Lichter

LMU Munich

Peter Quichen

Anna Friedl

Scripps Research Institute

Karin Monier

Kevin Sullivan

Supercomputing Center Karlsruhe

Rudolph Lohner

University Tübingen

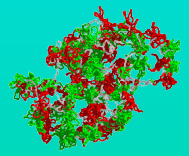
Markus Göker

Others from the DKFZ:
Monika Stöhr, Michael Stöhr, Andreas Hunziker, Angel Alonso

High-Performance Computing Center Stuttgart, University of Stuttgart; Supercomputing Center Karlsruhe, University of Karlsruhe; Computing Center, Deutsches Krebsforschungszentrum Heidelberg (DKFZ)

Bundesministerium für Forschung und Technology (BMFT) 01 KW 9602/2
(3D-Human Genome Study Group Heidelberg, German Human Genome Projekt)

Deutsches Krebsforschungszentrum (DKFZ), Erasmus MC Universität Heidelberg, Universität Karlsruhe, Universität Tübingen



From Sequence to Morphology

-

Long-Range Correlations in Complete Sequenced Genomes

Knoch, T. A.

Erasmus Medical Center, Erasmus University of Rotterdam, Rotterdam, The Netherlands,
16th December, 2004.

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, autofluorescent proteins, CFP, GFP, YFP, DsRed, fusionprotein, 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.

Fejes Tóth, K., **Knoch**, **T. A.**, Wachsmuth, M., Frank-Stöhr, M., Stöhr, M., Bacher, C. P., Müller, G. & Rippe, K. Trichostatin A induced histone acetylation causes decondensation of interphase chromatin. *J. Cell Science* 177, 4277-4287, 2004.

Ermler, S., Krunic, D., **Knoch, T. A.**, Moshir, S., Mai, S., Greulich-Bode, K. M. & Boukamp, P. Cell cycle-dependent 3D distribution of telomeres and telomere repeat-binding factor 2 (TRF2) in HaCaT and HaCaT-myc cells. *Europ. J. Cell Biol.* 83(11-12), 681-690, 2004.