DNA Sequence Patterns
a Successful Example of
Grid Computing in Genome Research
Building Virtual Super-Computers
for the Research Commons of e-Societies

Tobias A. Knoch
with
A. Abuseiris, R. de Graaf, M. Lesnussa, N. Kepper, Ad Emmen, & F. G. Grosveld

Biophysical Genomics & Erasmus Computing Grid
Dept. Cell Biology & Genetics, Erasmus Medical Center, Rotterdam, The Netherlands.
BioQuant Center & German Cancer Research Center (DKFZ), Heidelberg, Germany.

(c) 1996-2011 All Rights Reserved!
DNA Sequence Patterns
a Successful Example of
Grid Computing in Genome Research
and
Building Virtual Super-Computers
For the Research Commons of e-Societies

Tobias A. Knoch
with
A. Abuseiris, R. de Graaf, M. Lesnussa, N. Kepper, Ad Emmen, & F. G. Grosveld

Biophysical Genomics & Erasmus Computing Grid
Dept. Cell Biology & Genetics, Erasmus Medical Center, Rotterdam, The Netherlands.
BioQuant Center & German Cancer Research Center (DKFZ), Heidelberg, Germany.
The amount of information is growing exponentially with ever-new technologies emerging and is believed to be always at the limit. In contrast, huge resources are obviously available, which are underused in the IT sector, similar as e.g. in the renewable energy sector. Genome research is one of the boosting areas, which needs an extreme amount of IT resources to analyse the sequential organization of genomes, i.e. the relations between distant base pairs and regions within sequences, and its connection to the three-dimensional organization of genomes, which is still a largely unresolved problem. The underusage of resources as those accessible by grid with its fast turnover rates is very astonishing considering the barriers for further development put forward by the inability to satisfy the need for such resources. The phenomenon is a typical example of the *Inverse Tragedy of the Commons*, i.e. resources are underexploited in contrast to the unsustainable and destructing overexploitation in the *Classic Tragedy of the Commons*. An analysis of IT and the grid sector which attempts to share resources for better usage efficiency, reveals two challenges, which lead to the heart of the paradox: i) From a macro perspective all grid infrastructures involve not only mere technical solutions but also dominantly all of the autopoietic social sub-systems ranging from religion to policy. ii) On the micro level the individual players and their psychology and risk behavior are of major importance for acting within the macro autopoietic framework. Consequently, the challenges of grid implementation are similar to those of other pressing global issues e.g. climate protection. This is well described by extending the *Human Ecology* triangle to a rectangle: environment-individual-society-environment. By applying this extension of this classical field of interdisciplinary basic and applied research to the grid sector, i.e. by further extension to an *e-Human Grid Ecology* rational, the *Grid Inverse Tragedy of the Commons* can be understood and approached regarding the internalization challenge into e-Society and e-Life, from which then guidelines for the day-to-day management can be derived. This is of general importance for many complex fields and thus with similar paradoxes and challenges.

By using grid Long-range power-law correlations were found using correlation analysis on almost the entire observable scale of 132 completely sequenced chromosomes of $0.5\times10^9$ to $3.0\times10^7$ bp from Archaea, Bacteria, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*. The local correlation coefficients show a species specific multi-scaling behaviour: close to random correlations on the scale of a few base pairs, a first maximum from 40 to 3400 bp (for *Arabidopsis thaliana* and *Drosophila melanogaster* divided in two submaxima), and often a region of one or more second maxima from $10^5$ to $3\times10^5$ bp. Within this multi-scaling behaviour, an additional fine-structure is present and attributable to codon usage in all except the human sequences, where it is related to nucleosomal binding. Computer-generated random sequences assuming a block organization of genomes, the codon usage, and nucleosomal binding explain these results. Mutation by sequence reshuffling destroyed all correlations. Thus, the stability of correlations seems evolutionarily tightly controlled and connected to the spatial genome organization. In summary, genomes show a complex sequential organization related closely to their three-dimensional organization.

Consequently, grids can be established by solving the *Grid Inverse Tragedy of the Commons* using a *e-Human Grid Ecology* rational and indeed be used as e.g. in genome research for DNA sequence pattern analysis very successfully to determine for decades unresolved questions which demand very heavy IT support. Thus, indeed the solutions for the demand requirements in the research commons of e-Societies can be tackled successfully by such a systemic approach.
Dynamic and Hierarchical Genome Organization

The different organization levels of genomes bridge several orders of magnitude concerning space and time. How all of these organization levels connect to processes like gene regulation, replication, embryogenesis, or cancer development is still unclear?
Dynamic and Hierarchical Genome Organization

The different organization levels of genomes bridge several orders of magnitude concerning space and time. How all of these organization levels connect to processes like gene regulation, replication, embryogenesis, or cancer development is still unclear?
The Complexity of Cytogenetic Diagnostics

The process of cytogenetic analysis requires proper patient and sample analysis as well as a comprehensive evaluation of the results.
The Complexity of Cytogenetic Diagnostics

The process of cytogenetic analysis requires proper patient and sample analysis as well as a comprehensive evaluation of the results.

Better Cytogenetic Diagnostics is Based on Advanced Genome Understanding!
DNA Sequence Organization

Determination of the concentration fluctuation function $C(l)$ and its local slope the correlation coefficient $\delta(l)$ are an indication for the i) degree of long-rang scaling behaviour, ii) general multi-scaling, and iii) fine-structure features, which all are connected to all levels of genome organization and especially also to the three-dimensional genome architecture.

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

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

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

The largest desktop grid for the biomedical research and care sectors with now ~13 Tera FLOPS and a potential ~50 Tera FLOPS and ~20,000 desktops, at two city wide institutions: the Hogeschool Rotterdam and the Erasmus Medical Center.

Research:
- genomic and proteomic analysis
- epidemiology
- image analysis, e.g. Applied Molecular Imaging (AMI)

Education:
- training of the coming grid generation of IT specialists
- developing new concepts for grid computing

Diagnostics:
- clinical image and data analysis
- operation planning and operation support

Industry:
- brokerage of computing resources

~10 BioMedical User Groups

ECG - Centralized Office

Two Donor Organizations

~20,000 PC “Owners”, i.e. Local PC Donors.

Dedicated and Secured!
ECG Projects and Users

The projects and users currently run on the ECG are mainly basic research and application development in the biological and medical areas. Currently there are ~15 projects and ~10 users of the ECG with large expansion already in sight.

The ECG is one of the few grids on this level of complexity!

<table>
<thead>
<tr>
<th>Project</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Analysis</td>
<td>Dept. Medical Informatics</td>
</tr>
<tr>
<td>DNA Sequence Correlation</td>
<td>Dept. Cell Biology &amp; Genetics</td>
</tr>
<tr>
<td>Genome Duplication Analysis</td>
<td>Dept. Cell Biology &amp; Genetics</td>
</tr>
<tr>
<td>Epidemiology studies</td>
<td>Dept. Public Health</td>
</tr>
<tr>
<td>CT Image Analysis</td>
<td>Dept. Internal Medicine</td>
</tr>
<tr>
<td>Structural Genome Simulations</td>
<td>Dept. Cell Biology &amp; Genetics</td>
</tr>
<tr>
<td>Protein Motive Patterns</td>
<td>Dept. Urology</td>
</tr>
<tr>
<td>Mass Spectrometry Analysis</td>
<td>Center for Biomics</td>
</tr>
<tr>
<td>Chip &amp; Sequencing Analysis</td>
<td>Center for Genomics</td>
</tr>
<tr>
<td>Genome Wide Association Studies</td>
<td>Dept. Internal Medicine</td>
</tr>
<tr>
<td>Protein Structure Prediction</td>
<td>Dept. Cell Biology &amp; Genetics</td>
</tr>
<tr>
<td>Diffusion of Molecules</td>
<td>Dept. Pathology</td>
</tr>
<tr>
<td>Microscopic Image Analysis</td>
<td>Dept. Microscopy &amp; Cell Biology</td>
</tr>
</tbody>
</table>
ECG Basic Grid Structure

The ECG consists of the computer pools of the Erasmus MC and the Hogeschool Rotterdam. The client computers are controlled by the general ECG Emperor of the grid at EMC and a local master at HR. “Condor” is used as middleware since it is very well established and open-source.
ECG - Infrastructure

The infrastructure of the ECG consist of a dedicated architecture respecting client, user and management security and efficiency in respect to the fact that the ECG is one of the largest grids in the world with the highest possible degree of complexity.

The new ECG infrastructure guaranties the exploitation of the capabilities to the maximum!
MediGRID and Services@MediGRID operate the national biomedical research and care cluster-grid within the national German D-Grid initiative and integrate various disciplines, institutions, and states throughout Germany.

- Module Coordination:
  - coordination of the distributed office

- Module Resource Integration:
  - sharing of the integrated resources

- Module Middleware:
  - grid technical virtualization

- Module Ontology Tools:
  - ontology development for grid user projects

- Module BioMedical Informatics:
  - user projects in biomedical research

- Module Clinical Imaging:
  - user projects in clinical imaging

- Module Clinical Research:
  - user projects in general clinical research

- Module e-Science:
  - general research on e-grid science

Services@MediGRID:
- services towards MediGRID

Distributed - Office

~ 20 BioMedical User Groups

~ 20 D-Grid Communities

~ 100 D-Grid Donor Organizations (~10 MediGRID)

~10,000 Cluster Nodes

~ 2,000 Medical Secured

Dedicated and Secured!
Large-Scale Resource Sharing in IT: 

The Inverse Tragedy of the Commons

The grid phenomenon and its implications are similar complicated to the ecology/climate/environmental challenge!

The Tragedy of the Commons:

⇒ A resource belonging to all and being on limited demand is overexploited / destroyed by the users due to responsibility diffusion!

⇐ transforms into ⇒

The Inverse Tragedy of the Commons

A Resource belonging to all and being in affluent availability on limited demand is "underexploited" by potential users due to responsibility diffusion !!!!!

The grid challenge lies in the e-Social embedding of grid phenomenons:

• Micro-Social: the sharing attitude and socialization of the individual.
• Macro-Social: the organization culture of the embedding institution.
The social systems theory by Niklas Luhmann (1927-1998) based on the autopoietic concept of Humberto Maturana and Francisco Varela (1946-2001) is so far the most advanced social systems theory existing to describe the complexity of grid implementation.

The Social Sub-Systems Involved:

- Religion
- Education
- Science
- Art
- Economy
- Jurisdiction
- Policy

The e-Social challenge lies in the integration of sub-systems towards a working grid society:

- Micro-Sub-Systems: the sub-system stickiness of individuals.
- Macro-Sub-Systems: the integration of institutionalized sub-systems via soft interfaces.

The subsystems have their own code of communication and are separated from each other in a way blocking in principle a consistent integration although they form a society with all their contradictions !!!!!
Grid Psychology

From Individual to Cultural Risk Management

Grid implementation and social spread is carried first by individuals as with any meme introduced into society: in the focus of the transformation of society to e-Society stands - as always - the balance between potential risk and opportunity.

The Risk Psychology Matrix:

- Individual Security Perception & Risk Acceptance
- Knowledge-Based Security & Risk Acceptance
- Incidental Security Reaction Behaviour
- Legal and Political Security Scenarios
- Religious & Cultural Security Archetypi

The grid challenge lies in a unified concept addressing the psychology of grid:

- Micro-Risk-Management: the micro-risk in the perception the individual and its emotional well-being.
- Macro-Risk-Management: the macro-risks in the procedural and institutionalization in organizations.
Security of the ECG

One of the major issues concerning grids is to create a security level sufficient for all involved parties as we did for the ECG. The technology, culture and even the psychology of security have been and are treated very seriously by the ECG and are to the highest achievable level guarantying the integrity and privacy of the system and data!

**Security Technology**
- system hardening
- firewalling
- encrypted network communications
- intrusion detection monitoring
- logging of session
- auditing and testing of applications
- virus/trojan checking of in-/output
- authentication

**Security Culture**
- trained employees
- secure programming
- change management system
- constant check of procedure
- constant test and training of employees
- transparent management structure
- transparent infra structure
- NAN and ISO certification

The highest security level guarantying privacy is just good enough for us.
&
We still would like to have it better.
The success of grid is based on a sustainable grid ecology within the e-Society, i.e. the e-Human Ecology of Grid reaches a equilibrated space within the integration of grid psychology with autopoietic e-Social sub-systems. Human Ecology first evolved in Chicago in the 1920’s in the area of city development by Robert Park (1864-1944) and Ernest Burgess (1886-1966).

The Definition of e-Human "Grid" Ecology:

"Under e-Human "Grid" Ecology we understand the complete science of the relationships of grid to the surrounding environment to which we can count all conditions of existence in the widest sense."¹

¹ Haeckel, E., Generelle Morphology der Organismen, Berlin, Band 2, Allgemeine Entwicklungsgeschichte, p. 286, 1866.

² Haeckel, E., Natürliche Schöpfungsgeschichte, 9. Auflage, Berlin, p. 793, 1898

(c-Human "Grid" Ecology "is) …the relationship between grid and all other e-Social systems.

The solutions of the grid challenge on the operational layer are addressed by:

• Micro-Operationality: the participative integration of fundamental IT applications of major individual users complying with the psychology of grid in an e-Human Ecology manner.

• Macro-Operationality: the set-up of an open and sustainable management structure complying to all the autopoietic e-Social sub-systems in an e-Human Ecology manner.
The Curriculum of e-Human Ecology

Curricula of the canon and grammar of e-Human Ecology need to contain the further developed classic contents of the human ecology rectangle as well as the knowledge and understanding of the special aspects of e-Life and a-Life.

The classic human ecology triangle is transformed into the human ecology rectangle to complete the circle for creation and dependencies for a curriculum canon and grammar!
ECG - Management

To guaranty the integrity and efficiency of the ECG, a dedicated management structure has been put in place with corresponding checks and balances. The organization is made such, that also collaborations with other grids as e.g. the Dutch Big-Grid, the German nationwide MediGRID / D-Grid, or other European Grids as EGEE, EDGES can be exploited most efficiently.
Long-Range Correlations in DNA Sequences

Determination of the concentration fluctuation function $C(l)$ and its local slope the correlation coefficient $\delta(l)$ show that genomes show characteristic long-range correlations up to $10^8$ bp, i.e. in principle over their complete length. Beyond, they show a specific multi-scaling behaviour, as well as a characteristic fine-structure. All correlates with the 3D-architecture of genomes.

![Graphs and data plots](image-url)
Long-Range Correlations in DNA Sequences

Determination of the concentration fluctuation function $C(l)$ and its local slope the correlation coefficient $\delta(l)$ show that genomes show characteristic long-range correlations up to $10^8$ bp, i.e. in principle over their complete length. Beyond, they show a specific multi-scaling behaviour, as well as a characteristic fine-structure. All correlates with the 3D-architecture of genomes.
The concentration fluctuation function $C(l)$ and the correlation coefficient $\delta(l)$ show a characteristic long-range fine-structured multi-scaling behaviour, which is species specific. Tree construction thus leads to characteristic groups, which are similar to those suggested by classic phylogenetic trees. Thus, the sequence organization represents evolutionary lines.
Sequence Organization and Phylogeny

The concentration fluctuation function $C(l)$ and the correlation coefficient $\delta(l)$ show a characteristic long-range fine-structured multi-scaling behaviour, which is species specific. Tree construction thus leads to characteristic groups, which are similar to those suggested by classic phylogenetic trees. Thus, the sequence organization represents evolutionary lines.

![Graph showing correlation function $C(l)$ and correlation coefficient $\delta(l)$](image)

- The graph illustrates the correlation function $C(l)$ and the correlation coefficient $\delta(l)$ for various species, showing characteristic multi-scaling behaviour.
- The graphs are labeled with different species and classes, indicating distinct evolutionary lines.

**Species and Classes**
- Mean only Archaea
- Mean class A (without Archaea)
- Mean class A'
- Mean class A''
- Mean class B
- Mean Arabidopsis thaliana
- Mean Sacharomyces cerevisiae
- Mean Schizosacharomyces pombe
- Mean Drosophila melanogaster
- Mean Homo sapiens

**Window Size $l$ [bp]**
- The x-axis represents the window size in base pairs ($l$).
- The y-axis represents the correlation coefficient $\delta(l)$ and the correlation function $C(l)$.
The Happy End - Profits Sharing the Commons

Both the Erasmus Computing Grid and the MediGRID/Services@MediGRID examples show that the IT challenges mankind faces in the biomedical research and care sectors can be successfully approached by exploitation of the commons by e-Human “grid” Ecology means.

Simulation of the Organization of an entire Human Cell Nucleus

GLOBE 3D System-Biological Genome Information System
Acknowledgements

**Gridders:**
Anis Abuseiris (ECG R&D, systems, applications)  
Michael Lesnussa (ECG visualization)  
Nick Kepper (ECG parallel/MPI/CUDA)  
Rob de Graaf (ECG build/administration)

**Science:**
Markus Göker  
(phylogenies, DSMZ, Braunschweig)  
Rudolf Lohner  
(numeric, KIT, Karlsruhe)

**Collaborators:**
Ad Emmen (BOINC)  
Alexander Beck-Ratzka (BOINC)  
Martin Kühn (gaming infrastructures)

**Social Science:**
Volkmar Baumgärtner  
Julia Hampl  
Kurt Egger

**Steering Committee:**
Frank G. Grosveld, ErasmusMC  
Arthur van der Kamp, Erasmus MC  
Menno Kok, Erasmus MC

**Institutions and/or Funders:**
Erasmus Medical Center  
Hogeschool Rotterdam  
The Fraunhofer Society
The German MediGRID and Services@MediGRID  
The German D-Grid Initiatives  
The German Ministry for Science and Technology  
The Dutch Science Organization (NWO)  
The European EGEE Initiative  
The European EDGES Consortium  
The German Society for Human Ecology  
The International Society for Human Ecology  
The European Commission

© 1996-2011 All Rights Reserved
DNA Sequence Patterns

A Successful Example of Grid Computing in Genome Research and Building Virtual Super-Computers for the Research Commons of e-Societies

Knoch, T. A., Abuseiris, A., Lesnussa, M., Kepper, N., de Graaf, R. & Grosveld, F. G.

8th International Desktop Grid Foundation (IDGF) Workshop, Max Planck Institute for Gravitational Physics, Hannover, Germany, 17th August, 2011.

Abstract

The amount of information is growing exponentially with ever-new technologies emerging and is believed to be always at the limit. In contrast, huge resources are obviously available, which are underused in the IT sector, similar as e.g. in the renewable energy sector. Genome research is one of the boosting areas, which needs an extreme amount of IT resources to analyse the sequential organization of genomes, i.e. the relations between distant base pairs and regions within sequences, and its connection to the three-dimensional organization of genomes, which is still a largely unresolved problem.

The underusage of resources as those accessible by grid with its fast turnover rates is very astonishing considering the barriers for further development put forward by the inability to satisfy the need for such resources. The phenomenon is a typical example of the Inverse Tragedy of the Commons, i.e. resources are underexploited in contrast to the unsustainable and destructing overexploitation in the Classic Tragedy of the Commons. An analysis of IT and the grid sector which attempts to share resources for better usage efficiency, reveals two challenges, which lead to the heart of the paradox: i) From a macro perspective all grid infrastructures involve not only mere technical solutions but also dominantly all of the autopoietic social subsystems ranging from religion to policy. ii) On the micro level the individual players and their psychology and risk behaviour are of major importance for acting within the macro autopoietic framework. Consequently, the challenges of grid implementation are similar to those of other pressing global issues as e.g. climate protection. This is well described by extending the Human Ecology triangle to a rectangle: environment-individual-society-environment. By applying this extension of this classical field of interdisciplinary basic and applied research to the grid sector, i.e. by further extension to an e-Human Grid Ecology rational, the Grid Inverse Tragedy of the Commons can be understood and approached regarding the internalization challenge into e-Society and e-Life, from which then guidelines for the day-to-day management can be derived. This is of general importance for many complex fields and thus with similar paradoxes and challenges.

By using grid Long-range power-law correlations were found using correlation analysis on almost the entire observable scale of 132 completely sequenced chromosomes of 0.5x10^6 to 3.0x10^7 bp from Archaea, Bacteria, Arabidopsis thaliana, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens. The local correlation coefficients show a species specific multi-scaling behaviour: close to random correlations on the scale of a few base pairs, a first maximum from 40 to 3400 bp (for Arabidopsis thaliana and Drosophila melanogaster divided in two submaxima), and often a region of one or more second
maxima from $10^5$ to $3 \times 10^5$ bp. Within this multi-scaling behaviour, an additional fine-structure is present and attributable to codon usage in all except the human sequences, where it is related to nucleosomal binding. Computer-generated random sequences assuming a block organization of genomes, the codon usage, and nucleosomal binding explain these results. Mutation by sequence reshuffling destroyed all correlations. Thus, the stability of correlations seems evolutionarily tightly controlled and connected to the spatial genome organization. In summary, genomes show a complex sequential organization related closely to their three-dimensional organization.

Consequently, grids can be established by solving the Grid Inverse Tragedy of the Commons using a e-Human Grid Ecology rational and indeed be used as e.g. in genome research for DNA sequence pattern analysis very successfully to determine for decades unresolved questions which demand very heavy IT support. Thus, indeed the solutions for the demand requirements in the research commons of e-Societies can be tackled successfully by such a systemic approach.

**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 statistical mechanics, genomic uncertainty principle, genome function, genetics, gene regulation, replication, transcription, repair, homologous recombination, simultaneous co-transfection, cell division, mitosis, mitophase, 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, chromatin cross-linking, chromosome conformation capture (3C), selective high-resolution high-throughput chromosome interaction capture (T2C), 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, information browser, visual data base access, holistic viewing system, integrative data management, extreme visualization, three-dimensional virtual environment, virtual paper tool.

**Literature References**


Syndrome imprinting locus in human cell nuclei indicate maximum condensation changes in the few hundred nanometer range. *Differentiation* 76(1), 66-82, 2008.


