Inhomogeneous Sequences of Letters in DNA, Proteins and Languages

Jürgen Hauck,[1,*] and Klaus Mika[2]

[1] IFF Soft Matter Division, Research Center Jülich D–52425 Jülich, Germany
[2] IFF Theory III Division, Research Center Jülich D–52425 Jülich, Germany

Abstract
1D, 2D and 3D ordering of different species like the bases A, C, G and T of DNA sequences is characterized by the self-coordination numbers T_i of i = 1, 2, 3 first, second, etc. neighbors and concentrations x and y. The T_i or α_i values ($\alpha_i T_i^{\max} = T_i - (T_i^{\max} - T_i)$ x/y) of each species in a T_1, T_2, T_3 or α_1, α_2 structure map can be related with attractive or repulsive interactions of the species. Similar structure maps are obtained for a similar evolution process as outlined for different languages which were influenced by the Latin language. The DNA and protein sequences of enolase, secA or cytochrome enzyms can be ordered in different groups, which are supposed to be related by evolution.

Keywords: DNA sequences, languages, patterns, evolution
Abbreviations and notations:
T_i: self-coordination number of i = 1, 2, 3, … first, second, third, etc. neighbors,
α_i: short-range order parameter.

* correspondence J. Hauck; phone: 49-2461-614237; fax: 49-2461-612280; E-mail: j.hauck@fz-juelich.de

1 INTRODUCTION

The ordering of different species A, B, C, etc. can be characterized by the self-coordination numbers T_i with i = 1, 2, 3 of nearest, second and third neighbors and the ratio r = y/x of concentrations x of the analyzed species and y of all other species. The one-dimensional row of A and B with maximum coordination numbers $T_1^{\max}$ = $T_2^{\max}$ = $T_3^{\max}$ = 2 at distances a, $2a$ or $3a$ is the simplest example for ordering at composition A_xB_y with attractive or repulsive A–A interactions. The relation between structures and interactions can be outlined for A = ladies and B = gentlemen in a conference hall. Equal numbers of ladies and gentlemen (r = y/x = 1) can form three homogeneous structures with sequences AB,AB, etc. (opera house ordering), AABB,AABB, etc. (gay party) and $A_\infty B_\infty$ (moslem school). Each lady A (and each B) of the three structures has the self-coordination numbers 0 2 0 (opera), 1 0 1 (gay party) or 2 2 2 (moslem school) of other ladies. The complete segregation of ladies and gentlemen in a moslem school is achieved for infinite domains of ladies and gentlemen, where other T_i values at domain boundaries can be neglected. The T_i values or α_i values can be plotted in T_1, T_2, T_3 or α_1, α_2 structure maps with the three homogeneous structures at the corners of a triangle (first and second plot of Fig. 1).

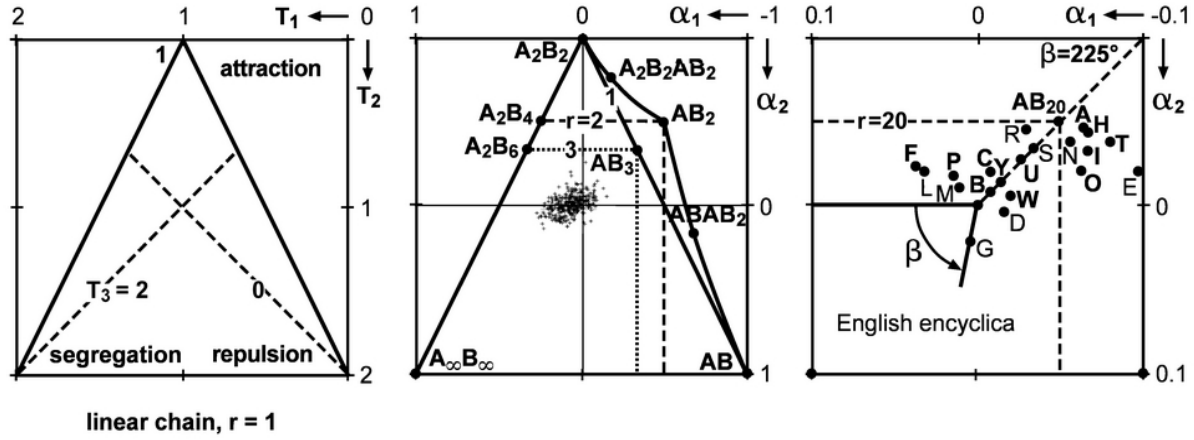

Figure 1. Structure map of linear chain with self-coordination numbers T_i of i = 1, 2 and 3 neighbors (first plot) or short-range order parameters α_1, α_2 as axes with α_1, α_2 values of 118 DNA sequences (+) and different periodic sequences of A and B (second plot). A section of this structure map with the α_1, α_2 values of each letter of the English translation of the encyclica *fides et ratio* is shown in the third plot. The letters V, K, Q, J, X, Z close to B position at β = 225 ^{0} are not shown.

The α_i values obtained from T_i values by $\alpha_i T_i^{\max} = T_i - (T_i^{\max} - T_i)$ x/y are more appropriate for evaluation of an increased number of species like A = mother, B = father, C = daughter and D = son of families or A = Adenine, C = Cytosine, G = Guanine and T = Thymine of DNA sequences (second plot of Fig. 1). The three homogeneous structures at the corners of a triangle can be related with repulsive A–A interactions (opera house), attractive A–A interactions (gay party) and segregation (moslem school). The different corners of the α_1, α_2 structure map were obtained numerically for increasing numbers n = x + y of seats with periodic boundary conditions. This can be compared with A and B sitting on a round table with n seats. The homogeneous AB,AB or AABB,AABB structures are obtained for n = 2 or n = 4 seats. The number of structures increases to 2^{n-1} as the round table is increased to n = x + y seats and the first site is occupied by A. A maximum of n = 24 seats were considered at y/x = 1 with $\binom{23}{11}$ = 1 352 078 structures. Few structures of the structure map like the $A_\infty B_\infty$ (moslem school) or $(A_2B)_\infty(AB_2)_\infty$ (triplet cluster) with T_i values 2 2 2 or 2/3 2/3 2 are derived by extrapolation. Tables with infinite size or separate tables for A (or A_2B) and B (or AB_2) are necessary, if all seats are occupied. Structures at the edges of the triangle are obtained by combination of corner structures like

ABA_2B_2, ABA_2B_2 (inhomogeneous boundary structures). This structure is inhomogeneous with different T_i values of the three A positions 0 1 2, 1 0 2 and 1 1 0, which are averaged to 2/3 2/3 4/3. Borders of the α_1, α_2 structure map connecting two structures with different r = y/x values, like the border between the A_2B_2 and AB_2 (or ABC) structure, are curved. At increased r values the α_1, α_2 structure map is reduced by the boundary at $\alpha_1 = \alpha_2 = 1/r$ ($r \geq 2$). A random distribution is obtained for $\alpha_i = 0$ (all r values). The entropy is increased for all structures, which are not on the border of the structure map (inhomogeneous metastable structures) to a maximum at $\alpha_i = 0$. Combinations of clusters A_nB_n at attractive interactions and isolated groups AB at repulsive interactions are inhomogeneous border structures for infinite n and inhomogeneous metastable structures for finite n like A_5B_5AB with T_i values 4/3 4/3 1. The situation is similar to mobbing of a single A who is separated by empty B seats from the A_5 cluster of people in a conference. These inhomogeneous structures, which are not on the border of the structure map, are supposed to be metastable. They should vary to stable border structures by an evolution process. Artificial metastable structures, which are obtained by computer-assisted design, are varied after some time.

Attractive or repulsive A–A forces or potentials can be calculated from homogeneous structures AB, AABB, AB_2 or $A_\infty B_\infty$ with maximum interactions, for inhomogeneous border structures and inhomogeneous metastable structures in the 1D row and the same method applied to 2D or 3D structures like the ordering of leaves (phyllotaxis), sunflower seeds, virus capsomeres or the coat patterns of zebras and giraffes [1,2]. Structures with maximum interactions can also be found by other forces as is outlined by the simple example of girls A and boys B in school. The attractive A–B interaction (or repulsive A–A interaction) is essential for the existence of the human being. Weak attractive or repulsive interactions could occur between the bases of DNA sequences or the amino acids of proteins containing atoms with different charges. The analyses of human languages and carols are probably somewhat easier to understand than the voices of animals like whales or birds. The evolution process of the Latin language could be studied in more detail, if old Italian, English of German texts would be analyzed by the present method.

2 RESULTS

Usually different T_i or α_i values are obtained for the different letters. The dots close to $\alpha_1 = \alpha_2 = 0$ in Fig. 1 (second plot) correspond to the α_i values of the four bases A (Adenine), T (Thymine), C (Cytosine) and G (Guanine) in 118 enzym DNA sequences of different species [3]. These values deviate from a random distribution. Some values are close to the borders at increased r values (not shown) of the α_1, α_2 structure map. The positions on the structure map are assumed to vary with time as the ordering process of ladies and gentlemen can vary with time. The T_i values of school children are usually varied from a point close to 2 2 2 ; (1) (moslem school of children at an early age) to a point 0 2 0 ; 1 (opera house) for teenagers. The DNA sequences can be ordered in different groups (Table 1). The ordering of the bases can be compared to the ordering of school children at an intermediate age. Most bases of the enzym enolase have a tendency to clustering (moslem school). A tendency to repulsion (opera house) is preferred by the bases of the chicken (*Gallus gallus*) and short human or rat DNA sections. The groups of the enzym secA are somewhat different with repulsive base interactions for the species *Halobacterium sp.* The evolution process of most species of one group is supposed to be similar. In that case the present groups can be compared with results obtained by other methods [4,5]. An increased number of clusters are suggested for the early evolution process [6].

Another example of evolution are the 26 letters of the Latin language compared to different other languages which were influenced by the Latin. Fig. 1 (third part) shows a section of the α_1, α_2 structure map for the English translation of the encyclica *fides et ratio* (vatican.de). Many letters are located on or close to the line $\alpha_1 = \alpha_2$ (or $T_1 = T_2$). These conditions are obtained for isolated A's. The $T_1 = T_2 = 0$ values correspond to negative α_i values. The letters S U Y V B etc. of the English language (values in brackets, Table 1) are usually in isolated positions. (The word 'mobbing' containing bb does not occur in the analyzed section of the encyclica with about 4000 letters.) These values at $\alpha_1 = \alpha_2$ are on a line with $\beta = 225^0$ of 2D polar coordinates (Fig. 1). The letters at increased β values R, C, P, M, L, F are also occurring as miniclusters like C

Table 1

Characterization of sequences of letters A, B, C, … in languages, proteins, carols or nucleic acids for example (GenBank of the National Center for Biotechnological Information) by the location of Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) bases on the α_1, α_2 structure map (Fig. 1) in 2D polar coordinates $\alpha = \sqrt{\alpha_1^2 + \alpha_2^2}$ and angle β ($\beta = 0^0$ or 90^0 for the values on the positive α_1 and α_2 axis). The letters written in bold are on the border of the structure map (inhomogeneous border structures). The F, G, etc. in carols can be FIS, GIS, etc. in some cases.

Species	Repulsive $[45^0]\ 45^0 < \beta < 225^0$	single (225^0)	attractive $225^0 < \beta$
Languages			
Latin encyclica	**VO**I**E**ST**A**RNPU	(□**QHGBXKWYZJ**)	MCD**LF**
Spanish encyclica	**VDOIEAST**□**MNU**	(C**PHBFQGYZXJKW**)	R**L**
Italian encyclica	V**OIEA**□**U**N	(R**HQJKWXY**)	**DC**MTSPL**FGZB**
Portuguese encyclica	**DMVOAEI**□**TN**SC	(**UPLQFHGBZXJKWY**)	R
French encyclica	M**GOEI**□**AU**T	(**QHVJBXYKWZ**)	R**D**SCN**PLF**
German encyclica	**G**HT**O**ED**AI**R	(□**UCWBZKVJQXY**)	NS**LFPM**
English encyclica	GDE**OWTIH**NA□	(S**UYVBKQJXZ**)	R**CP**ML**F**
Proteins			
Rice Enolase	A	(**LNQMFPSEHVTYC**)	**IGK**
Drosophila Enolase	[E]L**CTM**AD**PI**	(**NSGFQYRW**)	KV**H**
Ricinus Enolase	[L]**WNMRP**A	(V**TYQI**K**FC**)	GS**ED****H**
Yeast Enolase	[ET]AK**NPI**SV	(**RQY**L**MWC**)	G**FH**D
E.coli Enolase	[A]T**PDSL**K	(**FRYQWC**)	G**V**NIE**MH**
Staphylococcus Enolase	ATS**YP**E**N**DV**K**	(**FRQHWC**)	G**LIM**
Actinobacillus SecA	[HPI]V**CGF**ATK	(**W**)	LNEDS**YQM**R
Borrelia SecA	V**C**N**F**GAKT	(**QMPYHW**)	ELSDIR
Phormidium l. SecA	GFDTAE**N**	(**QSPMKHWC**)	RL**Y**IV
Streptomyces l. SecA	[FM]AQ**HVP**TGL	(**SYWC**)	**I**EDK**N**R
Mycobact. smeg. SecA	[PQ]AG**FTVS**E	(**YKHWC**)	LN**MI**DR
Human Cytochrome	[HRN]KVSD**AG**	(**PQMYWEC**)	**F**ILT
Camponotus f.Cytochr.	**DFL**I	(**AMYGNVEPR**)	**THKW**S
Silk GAGAG(SG(AG)$_n$)$_8$ SGAAGY	**G**A	(**SY**)	

Species	Repulsive $[45^0]\ 45^0 < \beta < 225^0$	single (225^0)	attractive $225^0 < \beta$
music notes			
O come, all ye faithful (Latin)	CAD**G**	(**HF**)	E
Silent Night (F. Gruber)	[GH]F	(**DE**)	**AC**
Joy to the World (G.F. Handel)	F**G**	( )	**HAC**DE
In Dulci Jubilo (German)	GC	(H)	**AD**F**E**
Hark, the Herald Angels Sing (F. Mendelssohn)	**C**	(A)	H**F**DGE
O Christmas Tree (German)		(**AFE**)	**D**CGH
Jingle Bells (J. Pierpont)		(**G**A	**F**DEHC

Table 1 continued

Species	rep.	sgl	att.	Species	rep.	sgl	att.
enolase				**secA**			
Drosophila melanogaster		( )	TGCA	*Actinobacillus actinomycet*		( )	CGTA
Loligo pealii		( )	TGCA	*Synechocystis sp.*		( )	CGTA
Sceloporous undulatus		( )	TGCA	*Cyanidium caldarium*		( )	CGTA
Eummeces inexpectatus		( )	TGCA	*Phormidium laminosum*		( )	CGTA
Oryza sativa		( )	TGCA	*Rickettsia prowazekii*		( )	CGTA
Pneumocystis carinii		( )	CGTA	*Bacillus subtilis* long		( )	GCTA
Saccharomyces cerevisiae		( )	CGTA	*Odontella sinensis*		( )	GCTA
Zymomonas mobilis		( )	CGTA	*Guillardia theta*		( )	GCTA
Entamoeba histolytica		( )	CGTA	*Staphylococcus carnosus*		( )	CGAT
Schistosoma japonicum		( )	CGTA	*Porphyra purpurea*		( )	CGAT
Python regius		( )	GCAT	*Mycobacterium tuberculosis*		(C)	GAT
Arabidopsis thaliana		( )	GCAT	*Heterosigma akashiwo*		( )	GCAT
Lycopersicon esculentum		( )	GCAT	*Mycoplasma pneumoniae*		( )	TGAC
Zea mays		( )	GCAT	*Borrelia burgdorferi*		( )	CTAG
Streptococcus therm.		( )	CGAT	*Aquifex aeolicus*		( )	GTAC
Nitrosomonas europea		( )	CGAT	*Mycoplasma genitalium*		( )	ATCG
Lycopersicon esculentum		( )	CGAT	*Prochlorothrix hollandica*		( )	ACGT
Ricinus communis		( )	GTCA	*Chlamydia pneumoniae*		( )	GATC
Chlamydomonas reinhardtii		( )	GTCA	*Streptomyces galbus*	A	( )	CTG
Candida albicans		( )	GCTA	*Prochloron didemni*	T	( )	GCA
Schizosaccharomyces pombe		( )	GCTA	*Mycoplasma capricolum*	C	( )	GAT
Rice m RNA		( )	CTGA	*Mycobacterium leprae*	C	( )	GAT
Neocallimastix frontalis		( )	TGAC	*Anabaena variabilis*	C	( )	GAT
Cunninghamella elegans		( )	TCGA	*Escherichia coli*	C	( )	GTA
Mesembryanth. crystall.		( )	GACT	*Haemophilus influenzae*	C	( )	ATG
Plasmodium falciparum		( )	GTAC	*Synechococcus sp.*	G	( )	CTA
Gluconobacter oxydans	T	( )	GCA	*Anacystis nidulans*	G	( )	CTA
Pelusios subniger	T	( )	GCA	*Bacillus firmus*	G	( )	CTA
Caiman crocodilus	T	( )	GCA	*Listeria monocytogenes*	G	( )	CAT
Alligator mississippiensis	T	( )	GCA	*Pisum sativum*	G	( )	CAT
Rattus norvegiens	T	( )	GCA	*Phormidium laminosum*	G	(C)	AT
Peking duck	T	( )	GCA	*Helicobacter pylori*	G	( )	TAC
Xenopus laevis	T	( )	GCA	*Spinacia oleracea*	[C]G	( )	AT
Trachemys scripta	T	(G)	CA	*Borrelia burgdorferi*	CG	( )	AT
Escherichia coli	T	( )	CGA	*Treponema pallidum*	CG	( )	AT
Mus musculus long	T	( )	CGA	*Staphylococcus aureus*	CG	( )	AT
Bos taurus	T	( )	CGA	*Arabidopsis thaliana*	CG	( )	AT
Mus musculus short	T	( )	AGC	*Bacillus subtilis* short	CG	( )	TA
Aspergillus oryzae	T	( )	AGC	*Zea mays*	CG	( )	TA
Homo sapiens long	T	( )	CAG	*Streptomyces coelicolor*	GT	( )	CA
Homo sapiens alpha short	T	(C)	AG	*Streptomyces lividans*	GT	( )	CA
Cladosporium herbarum	T	( )	GAC	*Streptomyces griseus*	GT	( )	CA
Staphylococcus aureus	C	( )	GAT	*Stylosanthes scabra*	AT	( )	CG
Alnus glutinosa	C	( )	GAT	*Antithamnion spec*	TA	( )	CG
Steptococcus intermedius	C	( )	GTA	*Pavlova luthgerii*	A	(G)	CT
Aspergillus oryzae	TC	( )	GA	*Caulobacter crescentus*	TG	( )	CA
Mastigamoeba balamuthi	TC	( )	GA	*Mycobacterium bovis*	GC	( )	TA
Fasciola hepatica	TG	( )	CA	*Vibrio alginolyticus*	GC	( )	TA
Sphenodon punctatus	TG	( )	CA	*Deinococcus radiodurans*	GC	( )	AT
Bacillus subtilis	GC	( )	TA	*Chlamydia trachomatis*	AGC	( )	T
Campylobacter fetus	CG	( )	TA	*Thermotoga maritima*	ACG	( )	T
Human m RNA very short	TCA	( )	G	*Mycobacterium smeg.*	GCT	( )	A
Rat neuron-specific short	CAT	( )	G	*Rhodobacter capsulatus*	TGC	( )	A
Gallus gallus	ATCG	( )		*Halobacterium sp.*	CGAT	( )	

Species	rep.	sgl	att.	Species	rep.	sgl	att.
cytochrome				**beta-globin [12]**			
Saccharomyces cerevisiae		( )	GCTA	*goat*		( )	TGCA
Bison bonasus	A	( )	CTG	*bovine*		( )	GTAC
Rat cytochrome	A	( )	CGT	*opossum*		(A)	TCG
Chimera cytochrome	A	( )	TCG	*chimpanzee*		(T)	GAC
Rhodobacter sphaeroides	T	( )	CGA	*human*		(GT)	AC
Streptomyces tendae	TA	( )	CG	*gorilla*		(GT)	AC
Capreolus capreolus	TA	( )	CG	*rat*	T	( )	GAC
Camponotus atriceps	CA	( )	TG	*mouse*	T	( )	GAC
Camponotus floridanus	CA	( )	TG	*gallus*	T	( )	GAC
Camponotus pennsylv.	CA	( )	GT	*rabbit*	TG	( )	AC
Human cytochrome	AT	( )	CG	*lemur*	[A] CTG	( )	

and R in ‘occurring’. The other letters at decreased β values are frequently occurring at alternating positions like I in ‘minicluster’. The letters G, D, E, N, S, R, M and L of the English language are either isolated, alternating or miniclusters (inhomogeneous metastable structures). The coexistence of isolated A’s and clusters of A’s (mobbing situation) with $T_1 = T_2$ at $\beta = 45^0$ are observed for some music notes or amino acids in proteins (values in squared brackets, Table 1). The sequence of four amino acids in silk to a structure close to an inhomogeneous border structure (last example of proteins in Table 1) can be related to the high strength. The deviation of Alanine (A) is essential for the elasticity of silk [7].

3 CONCLUSION

The relation between the type of ordering and attractive or repulsive interactions is outlined for the most simple example of the one-dimensional row. The homogeneous sequences AB, A_2B_2 of A and B are related to repulsive or attractive interactions. The A positions of boundary structures like ABA_2B_2 are inhomogeneous with two neighboring values $T_1 = 0$ or 1. Other inhomogeneous structures usually have more than two neighboring values like $T_1 = 0$, 1 and 2 for A_5B_5AB (inhomogeneous metastable structures).

Most letters of different languages and proteins are at the border of the structure map (inhomogeneous border structures written in bold in Table 1). The four bases of enzym DNA A, C, G and T are inhomogeneous metastable structures, which are not on the border of the structure map.

The human languages and DNA sequences of enolase of different species can be ordered to different groups. The similar sequences of Italian, Spanish and Latin languages or English and German languages can be explained by the evolution process. The voices of different animals like whales or birds could be analyzed by the same method as the music notes or the letters. Many animals have a single (A_∞) or two different sounds (AB), few like the voice of the cock or pigeon have the homogeneous AABB or ABC sequences. Homogeneous structures and border structures can also be found in coat patterns in an extended sense on surfaces of spheres (virus capsomeres, blackberries), cylinders (phyllotaxis, cylindrical viruses [1,8]) or other curved surfaces of animals. The capsomeres of the icosahedral microvirus have $T_1 = T_2 = 5$ and $T_3 = 1$ neighbors [2]. The capsomeres of most other viruses like simian virus 40, herpes virus or adenovirus are forming inhomogeneous border structures with $T_1 = 5$ and 6 [9,10]. The same applies to blackberries or sunflower seeds. The maximum density of the virus capsomeres on the surface of a sphere is essential to protect the virus genome. The cactus spines of some

cacti of the mamillaria family have a similar structure with T_1 = 5 or 6 neighbors for protection. The density of spines is decreased in *opuntiae cacti* with T_1 = 4 or columnar cacti with T_1 = 2. The coats of giraffes [11] contain white or yellow stripes with T_1 = 3 or 4 connections (T_1 = 1 or 2 for zebras, inhomogeneous border structures).

4 REFERENCES

(1) J. Hauck and K. Mika, Different Types of Order, *Cryst. Res. Technol.* **2003**, *38*, 831–846.
(2) J. Hauck and K. Mika, Attractive or Repulsive Interactions in Different Structure Types, *Progr. Solid State Chem.* **2003**, *31*, in press.
(3) J. Hauck, D. Henkel, and K. Mika, The evolution processes of DNA sequences, languages and carols, *Physica A* **2001**, *293*, 540–548.
(4) S. Guharay, B. R. Hunt, J. A. Yorke, and O. R. White, Correlations in DNA sequences across the three domains of life, *Physica D* **2000**, *146*, 388–396.
(5) T. Angata and A. Varki, Chemical Diversity in the Sialic Acids and Related α-Keto Acids: An Evolutionary Perspective, *Chem. Rev.* **2002**, *102*, 439–469.
(6) C. de Duve, *Ursprung des Lebens: Präbiotische Evolution und die Entstehung der Zelle (Blueprint for a cell: the Nature and Origin of Life)*, Spektrum, Akad. Verlag, Heidelberg, 1994.
(7) W. Hennig, *Genetik*, Springer, Berlin, 1998.
(8) I. Adler, D. Barabe, and R. V. Jean, A History of the Study of Phyllotaxis, *Annals of Botany* **1997**, *80*, 231–244.
(9) A. Klug and J. T. Finch, Structure of Viruses of the Papilloma-Polyoma Type. I. Human Wart Virus, *J. Mol. Biol.* **1965**, *11*, 403–423.
(10) T. S. Baker, N. H. Olson, and S. D. Fuller, Adding the Third Dimension to Virus Life Cycles: Three-Dimensional Reconstruction of Icosahedral Viruses from Cryo-Electron Micrographs, *Microbiology and Molecular Biology Reviews* **1999**, *63*, 862–922.
(11) P. Ball, *The Self-made Tapestry: Pattern Formation in Nature*, Oxford University Press, Oxford, 1999.
(12) C. Li and J. Wang, this conference

Biographies:
First author is senior research scientist with a Ph.D. degree in inorganic chemistry and crystallography from the University of Frankfurt (Germany).
Second author is senior research scientist with a Ph.D. degree in physics and mathematical modeling from the Technical University of Aachen (Germany).