Soft Matter

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Soft Matter
www.rsc.org/softmatter
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY
OF CHEMISTRY

www.rsc.org/softmatter

Statistical model of intra-chromosome contact maps

L.I. Nazarov,[a] **M.V. Tamm,**[a,e], **V.A. Avetisov**[b,e] **and S.K. Nechaev,***[c,d,e]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

A statistical model describing a fine structure of the intra-chromosome maps obtained by a genome-wide chromosome conformation capture method (Hi-C) is proposed. The model combines hierarchical chain folding with quenched heteropolymer structure of primary chromatin sequences. It is conjectured that observed Hi-C maps are statistical averages over many different ways of hierarchical genome folding. It is shown that the existence of quenched primary structure coupled with hierarchical folding induces full range of features observed in experimental Hi-C maps: hierarchical elements, chess-board intermittency and large-scale compartmentalization.

1 Introduction

Analysis of chromatin folding in human genome based on a genome-wide chromosome conformation capture method[1,2] provides a comprehensive information on spatial contacts between genomic parts and imposes essential restrictions on available 3D genome structures. The experimental Hi-C maps obtained for various organisms and tissues[2–8] (some examples are shown in the figures 1A-C) display very rich structure in a broad interval of scales. The researchers usually pay attention to the average contact probability, $\mathscr{P}(s)$, between two units of genome separated by a genomic distance, s, which decays in typical Hi-C maps approximately as $\mathscr{P}(s) \sim 1/s$ (see[2] and Fig.1D.

Apart from the averaged contact probability decay, the Hi-C maps show very rich fine structure behavior. The important features of the maps include (see, e.g. the inset in the Fig.1C): i) elements of a hierarchical structure on small scales[8], ii) chromosome compartmentalization on large scales, and iii) the chess-board intermittency in the color intensity[25].

Theoretical models of chromatin packing in the nucleus, which can possibly explain the observed behavior of intra-chromosome Hi-C contact maps, split roughly split into two groups. The first group of works relies on specific interactions within the chromatin, like loop or bridge formation,[9–15], while the second group aims to explain the chromatin structure in terms of large-scale topological interactions[2,16–23] based on so-called ”fractal” (or ”crumpled”) model of the polymer globule[24].

In view of presence of these competing theories, it seems very important to understand which experimentally observed phenomena can possibly be understood within each of them. In this paper we assume the crumpled globule approach and study how much the observed fine structure of Hi-C maps can be reproduced within this formalism. In particular, for the first time, to the best of our knowledge, we show that all the main features of the fine structure of Hi-C contact maps can be naturally obtained within the *heteropolymer* crumpled globule framework, where one can avoid specific biological details, and stick mainly to basic principles of statistical physics of disordered systems. Although the results presented below are mainly qualitative rather than quantitative, we believe that the presented approach can be refined to make our theory more system-specific. Our goal is to develop a ”bottom-up” theory: starting from the very basic physical principles, we construct a simplest possible model, study its behavior, and by comparing with real systems, get some insight into how the real biological system might work.

The crumpled globule is a state of a polymer chain which in a wide range of scales is self-similar and almost unknotted, forming a fractal space-filling-like structure. Both these properties, self-similarity and absence of knots, are essential for genome folding: fractal organization makes genome tightly packed in a broad range of scales, while the lack of knots ensures easy and independent opening and closing of genomic domains, necessary for transcription[16,18]. In a three-dimensional space such a tight packing results in a *space-filling* with the fractal dimension $D_f = D = 3$. The Hi-C contact probability, $P_{i,j}$, between two genomic units, i and j in a N-unit chain, depends on a combination of structural and energetic factors. Simple mean-field arguments (see, for example,[2]) demonstrate that in a fractal globule with $D_f = 3$ the *average* contact probability, $\mathscr{P}(s) = (N-s)^{-1}\sum_{i=0}^{N-s} P_{i,i+s}$, be-

[a] Physics Department, M.V. Lomonosov Moscow State University, 119992 Moscow, Russia
[b] N.N. Semenov Institute of Chemical Physics, RAS, 119991 Moscow, Russia
[c] Université Paris-Sud/CNRS, LPTMS, UMR8626, 91405 Orsay, France; E-mail: sergei.nechaev@gmail.com
[d] P.N. Lebedev Physical Institute, RAS, 119991 Moscow, Russia
[e] Department of Applied Mathematics, International Research University Higher School of Economics, 101000 Moscow, Russia

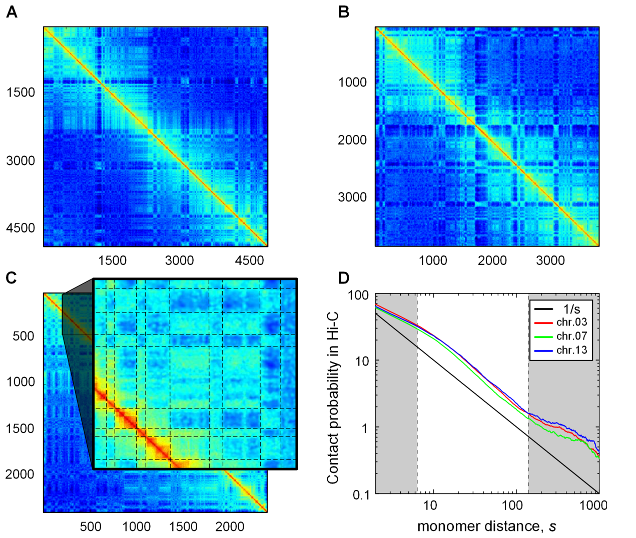


Fig. 1 (Color online) A-C: Samples of Hi-C maps (chromosomes 3 (A), 7 (B), and 13 (C), data provided by M. Imakaev), the color encodes the contact probability between genome fragments, each pixel corresponds to 40kb genome length; D: Contact probability decay in doubly-logarithmic coordinates for the same Hi-C maps (s is the genomic distance form the main diagonal of a map) compared with $1/s$ power law.

tween two units separated by the genomic distance $s = |i-j|$, decays as $\mathscr{P}(s) \sim s^{-1}$. It should be noted that recent numeric simulations[22,27], and more sophisticated arguments beyond the mean-field approximation[19,20], point out that the contact probability decays as $\mathscr{P}(s) \sim s^{-\gamma}$ with $\gamma \simeq 1.05 - 1.09$. This refinement can be easily taken into account in our approach, however in this paper we neglect it, since the slight deviation of γ from 1 goes beyond the accuracy of the simplest model considered in the present work.

Combining the assumption that chromatin can be considered as a heteropolymer chain with a quenched primary sequence[28], with the general hierarchical fractal globule folding mechanism, we are able to reproduce the large-scale chromosome compartmentalization, not assumed explicitly from the very beginning. To show the compatibility of the hierarchical folding of a crumpled globule with the fine structure of experimentally observed Hi-C maps we suggest a simple toy model based on the crumpled globule folding principles.

The paper is organized as follows. In the Section 2, to make a content of this paper as self-contained as possible, we highlight the basic concepts of the crumpled (fractal) globule formation and propose the new model of heterogenous Hi-C maps, which is analyzed in the Section 3. In the Section 4 we discuss the obtained results and speculate about possible developments of the proposed approach.

2 Heteropolymer crumpled globule

2.1 Basic principles of a crumpled globule formation

At high temperatures, i.e. in a good solvent, a polymer of N segments, each of length a, is a strongly fluctuating coil without a well-defined thermodynamic state. At temperatures below the θ-point (i.e. in a poor solvent), a polymer chain collapses into a weakly fluctuating, drop-like globule of size $R \sim aN^{1/3}$. In an ordinary globule, where the topological constraints are not taken into account, all subchains of length $l = as$, for $s \geq N^{2/3}$, appear as mutually entangled Gaussian coils, since volume interactions are screened in the melt, according to the Flory theorem[29]. However, in presence of topological constraints (especially for unknotted ring polymers), the globular state is essentially different. Forbidding knotting, one creates favorable conditions for the fractal globule formation with self-similar hierarchically folded crumples (folds), almost unknotted on all scales (see[24] for an original idea,[30] for a mathematical background, and[21,22] for recent extensive numeric investigations).

The collapse of a polymer into a crumpled globule state may be elucidated by the following imaginative hierarchical process. At the initial stage, there exists a certain length, $g^* = N_e/(a^6\rho^2)$ (N_e is the so-called the "entanglement length", and ρ is the globule density), such that the chain parts of the order of g^* collapse, constituting the ground-level (0-level) folds, which we denote as "units". Then, the chain segments, containing several consecutive units, collapse again, forming space-filling 1st-level folds; they in turn form 2nd-level folds, etc. The described process produces a hierarchy of folds (crumples), and ends when all g^*-link units are collected in a single (largest) fold – see the Fig.2a.

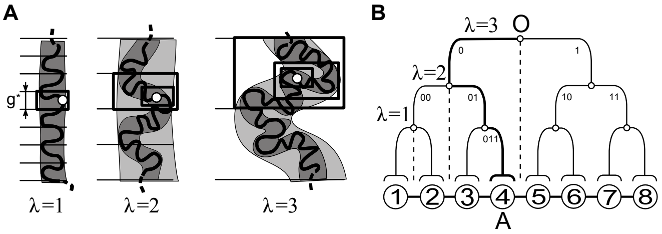


Fig. 2 (a) Schematic representation of few sequential stages of a hierarchical polymer folding and topological organization of hierarchically embedded crumples (folds); (b) Encoding of a position of a particular chain unit (the point A) in a set of folds by a descending path on a Cayley tree from a root (the point O) to the point A.

Recent extensive numeric simulations of collapsed unknotted polymer ring in a confining box[22], have demonstrated some differences between a non-equilibrium structure obtained immediately after a polymer collapse (called in[2] a

”fractal” globule) and equilibrium topologically constrained globular polymer ring (called in[24] a ”crumpled” globule). To make the content of the paper consistent with other works, we use below the name ”crumpled globule”.

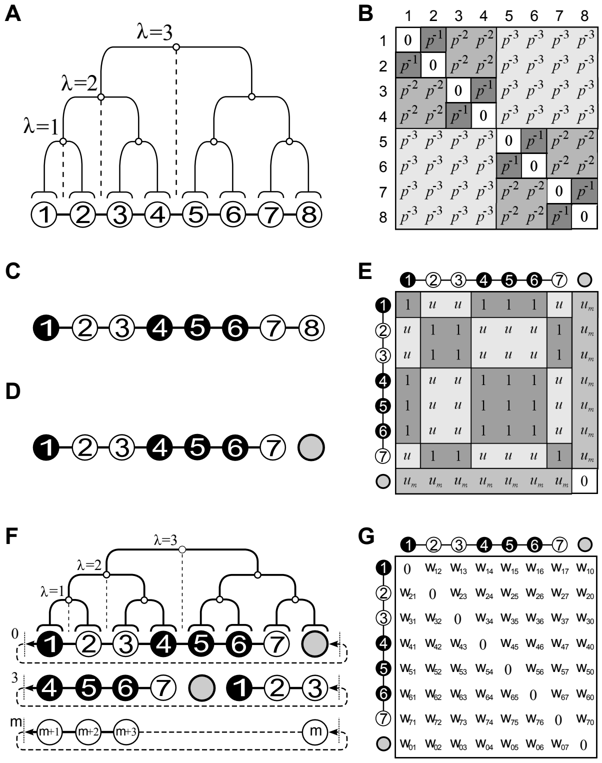
2.2 Model of a heteropolymer crumpling

In a hierarchically folded macromolecule position of each unit is characterized by a set of indices specifying to which particular 1st-level fold (embedded into particular 2nd-level fold, *etc*) this unit belongs. In a simplest case when each new fold on hierarchical level λ consists of $p = 2$ folds of preceding level $\lambda - 1$, the hierarchy of folds can be visualized by a Cayley tree, where the indexing is encoded by a path on the tree. Namely, associating binary numbers 0 and 1 to the left and to the right descending branches of a Cayley tree, respectively, as shown in Fig.2b, one can encode the ”coordinate” of some point A located in a terminal leaf in a binary sequence (011 in this particular case). This encoding provides a complete information how one can reach this point A from the root O of the tree and completely characterizes the position of a chain unit A in a set of crumples depicted in the Fig.2a.

Thus, the boundary nodes (leaves) of the Cayley tree constitute a ”space of states” for the chain units, and each subtree corresponds to a particular fold. The number, p, of downward Cayley tree branches ($p = 2$ in the Fig.2b), defines the number of λ-level folds embedded into one fold of the next hierarchical level, $\lambda + 1$ (we assume here that p is level-independent). It is convenient to choose the length of an elementary unit l such that $lp = g^*$, so the smallest fold consists of p elementary units. Then each λ-level crumple contains p^λ units.

The regularity of the hierarchy assumed above (one and the same value of p for all folds) is, of course, a rough oversimplification which we use here to construct the simplest possible toy model of hierarchical crumpling. Anyway, we believe that a regular tree model is still more realistic approximation to the description of a DNA hierarchical folding, than a random compactification without any reference to the hierarchy. Let us emphasize that the representation of the crumpled globule in terms of hierarchy of folds *a priori* does not capture the information about the volume interactions. However the excluded volume effect can be easily taken into account via the scaling dependence between the size of the crumple and the number of chain units in this crumple (see Eqs.(3) and (7) for details).

Now, after defining the set of folds as the hierarchically ordered tree, we need to specify how exactly, in which order, the chain itself fills these folds. By specifying this, we are able to combine the hierarchical geometry of folding with the linear geometry of the underlying chain. We assume that the chain fills the folds consequentially, but it has a freedom to choose the specific unit at which the folding begins. We can elucidate this by the following example shown in the Fig.3 for a periodic structure. Starting, for instance, with the fragment 1, we can unite fragments 1 and 2 in one fold of the level $\lambda = 1$, and do the same for the pairs [3,4], [5,6], [7,8],... On the level $\lambda = 2$ we thus have $\big[[1,2],[3,4]\big]$, $\big[[5,6],[7,8]\big]$, ... etc. However, when starting, say, from the fragment 4, we get hierarchical crumples [4,5], [6,7], [8,1], [2,3],... on the level $\lambda = 1$, $\big[[4,5],[6,7]\big]$, $\big[[8,1],[2,3]\big]$,... on the level $\lambda = 2$, etc. If all chain fragments are identical, all different possibilities of folding are equiprobable and this ambiguity leads to smearing of the hierarchical structure in the ensemble averages since there is no any preferred secondary structure. However, for a heteropolymer chain, the energies of direct contact interactions are fragment-dependent, and different foldings attain statistical weights depending on which specific units are in direct contact. This could remove the statistical degeneration of folding configurations in the ensemble averaging making some structure significantly preferred to others.



	1	2	3	4	5	6	7	8
1	0	p^{-1}	p^{-2}	p^{-2}	p^{-3}	p^{-3}	p^{-3}	p^{-3}
2	p^{-1}	0	p^{-2}	p^{-2}	p^{-3}	p^{-3}	p^{-3}	p^{-3}
3	p^{-2}	p^{-2}	0	p^{-1}	p^{-3}	p^{-3}	p^{-3}	p^{-3}
4	p^{-2}	p^{-2}	p^{-1}	0	p^{-3}	p^{-3}	p^{-3}	p^{-3}
5	p^{-3}	p^{-3}	p^{-3}	p^{-3}	0	p^{-1}	p^{-2}	p^{-2}
6	p^{-3}	p^{-3}	p^{-3}	p^{-3}	p^{-1}	0	p^{-2}	p^{-2}
7	p^{-3}	p^{-3}	p^{-3}	p^{-3}	p^{-2}	p^{-2}	0	p^{-1}
8	p^{-3}	p^{-3}	p^{-3}	p^{-3}	p^{-2}	p^{-2}	p^{-1}	0

	❶	②	③	❹	❺	❻	⑦	○
❶	1	u	u	1	1	1	u	u_m
②	u	1	1	u	u	u	1	u_m
③	u	1	1	u	u	u	1	u_m
❹	1	u	u	1	1	1	u	u_m
❺	1	u	u	1	1	1	u	u_m
❻	1	u	u	1	1	1	u	u_m
⑦	u	1	1	u	u	u	1	u_m
○	u_m	u_m	u_m	u_m	u_m	u_m	u_m	0

	❶	②	③	❹	❺	❻	⑦	○
❶	0	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}	W_{17}	W_{10}
②	W_{21}	0	W_{23}	W_{24}	W_{25}	W_{26}	W_{27}	W_{20}
③	W_{31}	W_{32}	0	W_{34}	W_{35}	W_{36}	W_{37}	W_{30}
❹	W_{41}	W_{42}	W_{43}	0	W_{45}	W_{46}	W_{47}	W_{40}
❺	W_{51}	W_{52}	W_{53}	W_{54}	0	W_{56}	W_{57}	W_{50}
❻	W_{61}	W_{62}	W_{63}	W_{64}	W_{65}	0	W_{67}	W_{60}
⑦	W_{71}	W_{72}	W_{73}	W_{74}	W_{75}	W_{76}	0	W_{70}
○	W_{01}	W_{02}	W_{03}	W_{04}	W_{05}	W_{06}	W_{07}	0

Fig. 3 A: Tree-like organization of crumples in a hierarchical folding; B: Hierarchical Parisi matrix of contact probabilities corresponding to space of states in A; C: Sample of quenched heteropolymer sequence with different types of units; D: The same as in C, but with the contacts between chain units and the ”outer space”, designated by open circles; F: Different translations are enumerated by the parameter m_f; G: Matrix of contact energies for a given primary structure; G: Composite weights of states obtained by superposition of matrices B and E – see (2).

Chromatin fiber (i.e., the complex of DNA and histone proteins) is a rather rigid object and can be presented as a sequence of renormalized monomers with an effective size of order of kilobases. Naïve guess would be to assume that the interaction constants between these renormalized monomers are quenched random variables with, say, a Gaussian distribution (compare[31]). However, the observed chessboard intermittency in the contact probabilities typical for Hi-C maps (see Fig.1) dictates a different point of view: it seems that the quenched interactions are clearly bimodal (multimodal in general), and it is more natural to model the heteropolymer by a quenched sequence of monomers of *several distinct types*. This assumption is supported by recent data on the variations in behavior of different chromatin types[28] and has been used lately in description of structure formation in bridge-stabilized models of chromatin[14,26]. In what follows we use the simplest option and model the chromatin by a sequence of two different units which we denote A and B. The chessboard intermittency of darker and lighter regions in contact maps is modelled by an Ising-type energy cost, $E_{i,j}$, associated with the spatial contact between two units i and j:

$$E_{i,j} = \begin{cases} -1, & \text{if } i \text{ and } j \text{ are of same type (A-A or B-B)} \\ -u, & \text{otherwise (A-B)} \end{cases} \tag{1}$$

where $0 \leq u \leq 1$ is the ratio of the energies of favorable and unfavorable contacts (henceforth we use the energy of a favorable contact as the energy unit).

The probability $P_{i,j}$ of a contact between ith and jth units of a hierarchically folded heteropolymer chain depends on the generation, $\lambda_{i,j}$, of the minimal common fold both these units belong to, and on the units types. We define the corresponding statistical weight of the ij contact as

$$w_{i,j}(\lambda) = e^{-\beta E_{i,j}} P_{i,j}^{\text{str}} + (1 - P_{i,j}^{\text{str}}), \tag{2}$$

where $E_{i,j}$ is the contact energy (1) between the units i and j, $P_{i,j}^{\text{str}}$ is an *a priori* (structural) contact probability between two units imposed by the hierarchy of folds, β is the inverse temperature, and the non-contact energy is assumed to be 0 by definition. The structural probability is a function of $\lambda_{i,j}$, and if folding is *space-filling*, then the mean-field expectation is

$$P_{i,j}^{\text{str}} \sim V^{-1}(\lambda_{i,j}) \sim p^{-\lambda_{i,j}}, \tag{3}$$

where $V(\lambda_{i,j})$ is the volume of the fold, which, for the space-filling folding, is proportional to the number of units in it. The proportionality coefficient in the r.h.s. of (3) can be absorbed into the definition of $E_{i,j}$. Thus, without the loss of generality, $P_{i,j}^{\text{str}} = p^{-\lambda_{i,j}}$. Note that this result relies on the conformation being space-filling: generally speaking, one expects $P_{i,j}^{\text{str}} = p^{-\alpha\lambda_{i,j}}$ with $\alpha < 1$ for significantly overlapping folds (this case is unphysical in the $N \to \infty$ limit as it does not respect the excluded volume constraint), and $\alpha > 1$ for a spongy globule with voids. In the Fig.3B,E we depict a Parisi-type hierarchical matrix P with elements $P_{i,j}^{\text{str}}$ and a matrix E with elements $E_{i,j}$ for some particular chain with a quenched monomer sequence.

To introduce averaging over realizations we proceed as follows. Take a polymer chain with a given quenched sequence of units and consider all possible ways to fold it into hierarchical structures. Since in our model the geometry of the tree of folds is fixed, and the chain fills all folds sequentially, there is only one possible way to alter the folding structure from one realization to the other, that is to change the position of the first elementary unit at the tree boundary, see Fig.3F. This change of the starting position induces a cyclic shift of monomers along the boundary: $i \to i + m \,(\operatorname{mod} p^{\lambda_{\max}})$, $i = 1, 2, ..., N$, where N is the chain length. The parameter defining a particular folding configuration, is the shift m. In the Fig.2F the samples corresponding to $m = 0$ and $m = 3$ are shown.

Now, one can self-consistently define the weights of particular hierarchical foldings. Assume that all contacts within a given folding are formed independently (i.e. all correlations in contact formations are already encoded in the underlying tree structure). The total folding weight can be written as a product of individual weights, $W(m) = \prod_{i,j=1}^{p^{\lambda_{\max}}} w_{i,j}(\lambda|m)$. The weights $w_{i,j}$ $(i, j = 1..N < p^{\lambda_{\max}})$ are given by (2). One should make an additional assumption about the weights of the contacts between chain folds and the "outer space", i.e. of the chain parts surrounded by other molecules designated by open circles in the Fig.3D. Since the chain folding happens in a nucleus within a crowded environment, one assumes that these open circles are effectively filled by the units of other chromosomes. To account for that, we introduce a mean-field interaction between units of the chain under consideration and the "average" units of outer chains similar to (2), but here $j > N$:

$$E_{i,j} = \begin{cases} -q - u(1-q), & \text{if } i \text{ is of type A} \\ -(1-q) - qu, & \text{if } i \text{ is of type B} \end{cases} \tag{4}$$

and q is an average fraction of monomers of type A. The total partition function accounting for all folds, reads now

$$Z = \sum_{m=0}^{p^{\lambda_{\max}}-1} W(m) \tag{5}$$

The probability for each pair of units, i and j is, as usual in equilibrium statistical mechanics,[32]

$$P_{i,j} = \sum_{m=0}^{N-1} \frac{W(m)}{Z} \times \frac{e^{-\beta E_{i,j}} P_{i,j}^{\text{str}}(m)}{e^{-\beta E_{i,j}} P_{i,j}^{\text{str}}(m) + (1 - P_{i,j}^{\text{str}}(m))}, \tag{6}$$

where the first term defines the thermodynamic probability of a particular hierarchical folding realization, and the second term explicitly encounters for the contact probability of this particular folding. The values of $P_{i,j}$ given by (6) are the contact probabilities that should be compared with the results of Hi-C measurements for intra-chromosome contact maps.

In the high-temperature limit, i.e. for $\beta \to 0$, one has $w_{ij} = 1$ as it follows from (2), and all $P_{i,j}$ are just the averages of $P_{i,j}^{\rm str}$ over cyclic permutations of indices along the Cayley tree boundary. Being averaged, all $P_{i,j}$ depend only on $s = |i-j|$, and in the limit $N \gg 1$ they are $\mathscr{P}(s = |i-j|) \sim s^{-1}$. To get this scaling, consider the values of $\mathscr{P}(s)$ for $s = p^m$, $m = 0, 1, 2....$ Let also N be the power of p, i.e. assume that $N = p^M$. Then

$$\mathscr{P}(p^m) = \frac{N}{N-1}\sum_{i=1}^{M-m}\frac{(1-p)}{p^i}\frac{1}{p^{(i+m)}} \simeq C\,p^{-m} \qquad (7)$$

where $C = [p^2(1+p)]^{-1}$. The last equation is valid for $(M-m) \gg 1$. Since $p^m = s$, we get $\mathscr{P}(s) \sim s^{-1}$, which is the consequence of the ”space filling” supposition. Namely, any crumple contains only nearest-neighboring monomers along the chain. This guarantees that in the volume occupied by the crumple there is no space to place monomers from other parts of the chain. This is the consequence of the space-filling supposition: the crumple of hierarchical level m has the volume, which is p^m times the volume of initial unit, and simultaneously, the crumple contains exactly $s = p^m$ neighboring along the chain initial units. This assumption reproduces (at least for $\beta \to 0$) average the dependence of the contact probability, $\mathscr{P}(s) \sim s^{-1}$. Note, that to reproduce the scaling $\mathscr{P}(s) \sim s^{-\gamma}$ with $\gamma \simeq 1.05 - 1.09$ for the averaged contact probability, measured in real Hi-C contact maps and numerical simulations for fractal globules (see the Introduction for the references) one should tweak with (3) by substituting $\lambda_{i,j} \to \gamma\lambda_{i,j}$, which would suggest that the fold of hierarchical level λ contains more than p^λ elementary units.

Summing up, the input of our model consists of a heteropolymer primary monomer sequence, and three numerical parameters: (i) the number of subfolds, p, embedded in each fold of the hierarchy (this parameter, though being quantitatively important, does not influence the qualitative appearance of the resulting structure of Hi-C maps), (ii) the ratio, u, of contact energies AB to that of AA/BB, and (iii) the inverse temperature, β, which essentially regulates the uniqueness of folding: for $\beta \to 0$ all foldings are equivalent and equally contribute to the resulting probability, while for $\beta \to \infty$ the folding with the lowest energy give the dominant contribution to contact probabilities.

3 The results

In the Fig.3 and Fig.4 examples of contact maps generated by our model are shown. To demonstrate the influence of a monomer composition on the contact probability, we have generated random Markovian heteropolymer sequences of $N = 150$ monomers with varying average block lengths (average lengths of blocks A and B are equal, thus the average fraction of units ”A” is $q = 0.5$). The parameters $u = 0.25$, $\beta = 0.25$, and $p = 2$ are the same for all three sequences. One sees that the resulting contact maps are essentially sequence-dependent, while the averaged contact probability decay plotted in the Fig.3D as a function of the genomic distance, s, still approximately follows the $1/s$ power law for all selected sequences.

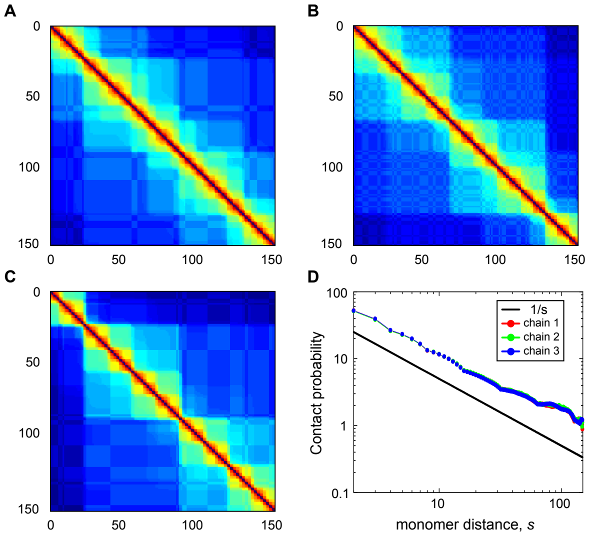


Fig. 4 (Color online) A-C: Dependencies of contact probability of the hierarchical fractal globule on the primary sequence ($N = 150$) for different realizations of primary sequences and fixed $u = 0.25$, $\beta = 0.25$ and $p = 2$; the average length of blocks is 10 for (A), 3.33 for (B) and 12.5 for (C); D: Averaged contact probability for sequences in A-C compared to the $1/s$ plot.

In the Fig.4A,B and C we demonstrate the influence of β on the contact probability maps for a fixed heteropolymer chain of length $N = 150$ with average block length 12.5 and $u = 0.25$. One sees the sequential degradation of the block-hierarchical structure with increasing temperature. The Fig.4D shows that the averaged contact probability is almost temperature-independent.

It should be emphasized that the experimental Hi-C contact map for a particular chromosome (as shown in the Fig.1) represents the averaged contact map over the ensemble of different (about millions) chain foldings. By inspecting experimental Hi-C maps, corresponding to different chromosomes, one can note that the hierarchical structure is not always prominent: for some chromosomes it is clearly seen, while it is smeared for others. We suggest that the manifestation of hierarchical structure is deeply connected (within the framework of the fractal globule concept) with the uniqueness of chro-

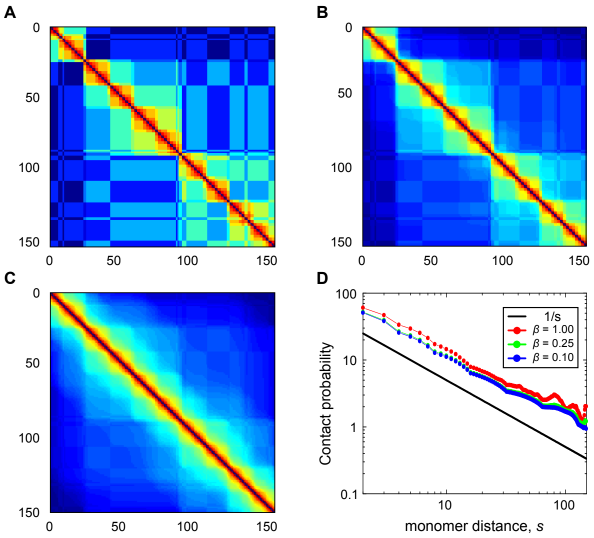


Fig. 5 (Color online) A-C: Dependencies of contact probability of the hierarchical fractal globule on the inverse temperature β for fixed initial sequence of $N = 150$ monomers, with $u = 0.25$ and $p = 2$, A: $\beta = 1$, B: $\beta = 0.25$, C: $\beta = 0.1$; D: Comparison with $1/s$ plot.

matin folding: if for a given quenched primary sequence, the hierarchies of folds are arranged similarly in different folding realizations in ensemble, then the block-hierarchical structure (typical for each realization), is clearly seen, while if the folds are arranged differently from one realization to the other, the hierarchical structure is smeared out due to the degeneration of corresponding Boltzman weights of particular foldings), however the $1/s$-decay of an average contact probability holds. In particular, we have demonstrated that combination of the hierarchical heteropolymer structure of a single folding with averaging over different foldings reproduces the typical behavior of experimentally observed Hi-C maps.

We conjecture that smearing of contact maps of *individual* chain foldings plays the same role as an overlap of replicas in statistical theory of random heteropolymers with quenched primary sequences[33]. To make this connection more profound, recall[34] that in a spin glass the overlap of two pure states α and β is characterized by the matrix $q^{\alpha\beta} = N^{-1}\sum_{i=1}^{N} m_i^{\alpha} m_i^{\beta}$, where m_i^{α} (or m_i^{β}) is the magnetization of the state α (or β) at a point i. The probability, $Q(q)$, to have an overlap q for any pair of states α and β in the system, is $Q(q) = \sum_{\alpha,\beta} Q_{\alpha} Q_{\beta}\, \delta\left(q - q^{\alpha\beta}\right)$, where Q_{α} is the probability of the state α. Along the same lines, it is natural to introduce an "overlap", q, of two different adjacency matrices (individual contact maps), P_{α} (consisting of elements $P_{i,j}^{\alpha}$) and P_{β} (consisting of elements $P_{i,j}^{\beta}$) for hierarchically folded chain with fixed primary sequence in the ensemble of M such matrices*. The overlap q can be defined as a "scalar product" of a pair of matrices, averaged over the ensemble, namely, $q = M^{-1}\sum_{\alpha\neq\beta}\left\langle \mathsf{P}_{\alpha}\mathsf{P}_{\beta}\right\rangle$. The "scalar matrix product", $\left\langle \mathsf{P}_{\alpha}\mathsf{P}_{\beta}\right\rangle$, can be constructed using the so-called "singular value decomposition"[35], meaning that the matrix P_{β} is evaluated in the basis of the matrix P_{α}. Certainly, constructing of such an overlap, q is accessible still only in the numerical simulations on model systems. The corresponding work is in progress.

4 Discussion and perspectives

Fine structure of contact probabilities observed in the model is reminiscent of that obtained in the experimental Hi-C maps – compare Figs. 3 and 4 with Fig.1. Note that the compartmentalization in our maps (i.e., large-scale block structure) is induced by the quenched primary sequence on a much smaller length scale (contrary, e.g. to[14] where compartmentalization is dictated by the heteropolymer structure of the chain on the very same lengthscale). Indeed, the configuration of large-scale blocks is highly disorder-dependent. In absence of any disorder all folding configurations have equal Boltzmann weights (i.e. are degenerated), so the average contact probability decays gradually as $1/s$ with genomic distance s. Certainly, the model proposed here is a mere caricature of a real situation: the number of possible ways of folding in our model grows linearly with the chain length, while it is bound to be exponential in real life, where the hierarchical folding mechanism accounts for intrinsic randomness. However, we believe that the main result, i.e. the emergence of fine structure due to the interference of many folding configurations with different statistical weights, will persist.

To summarize, in this paper we have proposed a statistical model which reproduces principal features of experimentally observed Hi-C maps. We took into consideration the heteropolymer structure and combined it with the hypothesis of hierarchical chromatin folding. We tried to avoid the specific biological details, sticking mainly to basic principles of statistical physics of disordered systems. Such a description, being less informative for concrete biological systems, allows us to conjecture the generic mechanism behind the fine structure of Hi-C maps and could be considered as a complimentary to the probabilistic refinement of Hi-C experiments developed recently in[36].

We have assumed that each single chromosome conformation is hierarchically folded with its own contact map. Considering the ensemble of different chain foldings for a given quenched primary sequence, and putting all particular contact

* Let us emphasize that the contact maps obtained in Hi-C experiments are the averages over ensemble of $\sim 10^7$ different folding realizations for fixed primary sequence.

maps on top of each other, we predict two typical scenario: (i) if the primary heteropolymer sequence dictates the typical (unique) folding, then the hierarchical structure is clearly seen in experimental Hi-C contact maps, however (ii) if the primary sequence does not dictate the typical folding, the block-hierarchical structure of experimental Hi-C contact maps is smeared out. We should emphasize the crucial importance of quenched disorder in primary sequence: just due to the presence of disorder, different ways of chromatin folding in our model have different energies and different statistical weights. In absence of disorder (i.e. for a homopolymer chain) or at high temperatures all possible folded structures would have the same energy, and after averaging over all states the contact maps will be plain gradient maps with contact probability depending only on the genomic distance between the chromatin fragments.

Besides, we see that in our model the largest compartments get smeared the least, because the shift in the position of largest fold corresponds to the largest energy barrier. We believe that the hierarchical compartmentalization of chromosomes into large domains, widely observed in experiments is facilitated by a collective effect of many *similar* microscopic monomer-monomer interactions of heteropolymer primary structure folded as a fractal globule. In connection with that, it would be very interesting to check if there is a correlation between the chromosome function and the manifestation of block-hierarchical structure of Hi-C intra-chromosome contact maps.

The authors are grateful to A. Grosberg, M. Imakaev, L. Mirny, A. Mironov and J. Mozziconacci for stimulating discussions. This work was initiated during the visits of M.T. and L.N. to LPTMS (Universite Paris-Sud), and LPTMC (Universite Paris 6), the authors appreciate the hospitality of their hosts as well as the financial support from the IRSES project FP7-PEOPLE-2010-IRSES 269139 DCP-PhysBio. This work was partially supported by the RFBR grant 14-03-00825. M.T., V.A. and S.N. acknowledge the financial support of the Higher School of Economics program for Basic Research.

References

1 J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, Science **295** 1306 (2002)

2 E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, *et al*, Science **326** 289 (2009)

3 J.E. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren, Nature **485** 376 (2012)

4 T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, Cell **148** 458 (2012)

5 Y. Zhang, R.P. McCord, Y.-J. Ho, B.R. Lajoie, D.G. Hildebrand, A.C. Simon, M.S.Becker, F.W. Alt, and J. Dekker, Cell **148** 908 (2012)

6 S. Sofueva, E. Yaffe, W.-C. Chan, D. Georgopoulou, M.V. Rudan, H. Mira-Bontenbal, S.M. Pollard, G.P. Schroth, A. Tanay, and S. Hadjur. The EMBO Journal, advance online publication (2013) doi:10.1038/emboj.2013.237

7 T.B.K. Le, M.V. Imakaev, L.A. Mirny, and M.T. Laub, Science **342** 731 (2013)

8 J. Dekker, M.A. Marti-Renom, and L.A. Mirny, Nature Reviews Genetics **14** 390 (2013)

9 R.K. Sachs, G. van der Engh, B. Trask, H. Yokota, and J.E. Hearst, Proc. Nat. Acad. sci., **92**, 2710 (1995)

10 C. Münkel, and J. Langowski, Phys. Rev. E, **57**, 5888 (1998)

11 J. Ostashevsky, Mol. Biol. of the Cell, **9** 3031 (1998)

12 J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heerman, R. van Driel, and S. Goetze, Proc. Nat. Acad. Sci., **106**, 3812 (2009)

13 B.V.S. Iyer, and G. Arya, Phys. Rev. E, **86**, 011911 (2012)

14 M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, Proc. Nat. Acad. Sci., **109**, 16173 (2012)

15 C.C. Fritsch, J. Longowski, Chromosome Res., **19**, 63 (2011)

16 A.Y. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23** 373 (1993)

17 A. Rosa, R. Everaers, PLoS Computational Biology, **4**: e1000153 (2008).

18 L.A. Mirny, Cromosome Res. **19** 37 (2011)

19 J.D. Halverson, J. Smrek, K. Kremer, and A. Yu. Grosberg, Rep. Progr. Phys., **77**, 022601 (2014)

20 A.Yu. Grosberg, Soft Matter, **10**, 560 (2014)

21 A. Rosa, R. Everaers, Phys. Rev. Letters, **112**, 118302 (2014)

22 M. Imakaev, K. Tchourine, S. Nechaev, and L. Mirny, arXiv:1404.0763

23 M. Tamm, L. Nazarov, A. Gavrilov, and A. Chertovich, arXiv:1404.2558

24 A.Yu. Grosberg, S.K. Nechaev, and E.I. Shakhnovich, J. de Physique **49** 2095 (1988)

25 A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci, BMC Genomics **13** 436 (2012)

26 M.Barbieri, A. Scialdone, A. Gamba, A. Pombo, and M. Nicodemi, Soft Matter, **9**, 8631 (2013).

27 J.D. Halverson, W.B. Lee, G.S. Grest, A.Y. Grosberg, and K. Kremer, J. Chem. Phys., **134**, 204905 (2011).

28 G.J. Filion, J.G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L.D. Ward, W. Brugman, I. de Castro Gene-

bra de Jesus, R.M. Kerkhoven, H.J. Bussemaker, and B. van Steensel, Cell, **143**, 212 (2010).

29 P.-G. de Gennes, Scaling Concepts in Polymer Physics, Cornell University Press, NY, 1979.

30 S. Nechaev and O. Vasilyev, Thermodynamics and topology of disordered knots: Correlations in trivial lattice knot diagrams, in ”Physical and Numerical Models in Knot Theory”, chapter 22, pp. 421-472, *Series on Knots and Everything*, (WSPC: Singapore, 2005)

31 E.I. Shakhnovich, and A.M. Gutin, J. Phys. A, **22**, 1647 (1989).

32 L.D. Landau, E.M. Lifshitz, Statistical Physics, Part 1 (Elsevier: Oxford, 1980)

33 C.D. Sfatos, A.M. Gutin, and E.I. Shakhnovich, Phys. Rev. E **48** 465 (1993)

34 M. Mezard, G. Parisi, M. Virasoro, *Spin glass theory and beyond* (World Scientific: Singapore, 1987)

35 G.H. Golub and C.F. Van Loan, *Matrix Computations* (4th ed.) (Johns Hopkins Studies in Math. Sciences, 1996)

36 E. Yaffe and A. Tanay, Nature Genetics **43** 1059 (2011)