Available online at www.sciencedirect.com
ScienceDirect
Physics of Life Reviews 8 (2011) 39–50
PHYSICS of LIFE reviews
www.elsevier.com/locate/plrev

Review

Cracking the chromatin code: Precise rule of nucleosome positioning

Edward N. Trifonov [a,b,*]

[a] *Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel*
[b] *Department of Functional Genomics and Proteomics, Faculty of Science, Masaryk University, Kotlarska 2, CZ-61137 Brno, Czech Republic*

Received 20 December 2010; received in revised form 12 January 2011; accepted 14 January 2011

Available online 19 January 2011

Communicated by M. Frank-Kamenetskii

Abstract

Various aspects of packaging DNA in eukaryotic cells are outlined in physical rather than biological terms. The informational and physical nature of packaging instructions encoded in DNA sequences is discussed with the emphasis on signal processing difficulties – very low signal-to-noise ratio and high degeneracy of the nucleosome positioning signal. As the author has been contributing to the field from its very onset in 1980, the review is mostly focused at the works of the author and his colleagues. The leading concept of the overview is the role of deformational properties of DNA in the nucleosome positioning. The target of the studies is to derive the DNA bendability matrix describing where along the DNA various dinucleotide elements should be positioned, to facilitate its bending in the nucleosome. Three different approaches are described leading to derivation of the DNA deformability sequence pattern, which is a simplified linear presentation of the bendability matrix. All three approaches converge to the same unique sequence motif CGRAAATTTYCG or, in binary form, YRRRRRYYYYYR, both representing the chromatin code.

Keywords: Ducleotide periodicity; DNA bendability; Matrix of bendability; Nucleosome sequence pattern; Nucleosome mapping; Signal processing

1. Basics of chromatin structure

1.1. Introduction

Centimeters long DNA molecules of higher organisms are squeezed in the chromosomes of the size of microns, with 10^5–10^6-fold compaction. If the semi-rigid chain of 10 cm long DNA (one chromosome) is freely suspended in water solution, it would make a random coil of the size $10^{3.5}$ times smaller than its length, due to its natural flexibility (statistical segment 100 nanometers [1]). Nature, thus, should have taken care of only additional compaction of about two orders. It is, however, not only matter of compaction. The problem is to fold it in such a way that in the course of cell division rather quick unfolding would be allowed, without entanglement. It then folds back to the original compact state. The most compact metaphase chromosomes, right before cell division, appear in cytological slides as

* Address for correspondence: Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel. Tel.: +972 4 828 8096; fax: +972 4 824 6554.
E-mail address: trifonov@research.haifa.ac.il.

1571-0645/$ – see front matter
doi:10.1016/j.plrev.2011.01.004

sets identical for all cells of given species, with bands of stain making very specific and different patterns along each of the 5 to 100 chromosomes, depending on species. That is, DNA of, say, human chromosome 9 is folded in a unique way characteristic for that chromosome. The instructions about the details of the folding are somehow stored in DNA itself, in its nucleotide sequence. The code that translates the linear sequence of DNA bases into its 3D trajectory, that is, physical and geometrical nature of the DNA folding, and how it is expressed in the sequence form, is the subject of this review.

1.2. Two obvious modes of organized folding

One solution of the problem would be the folding scheme used for shroud lines of a parachute. Here the lines are folded in a zig-zag (side-by-side) manner, quick to unfold to full length without twisting or knotting. One could also wind the rope in a cylindrical spring, solenoid. In this case pulling the ends would not cause an entanglement but, obviously, a significant twist will be inflicted to the fully extended rope. One also could imagine some designs with compensating windings of opposite sign. As long as DNA is concerned Nature utilizes both folding schemes, zig-zag (e.g. [2]) and winding modes [3], although it is not clear yet, what is the actual large scale arrangement (higher order structure) that guarantees quick reversible folding/unfolding. Remarkably, DNA does suffer some degree of the entanglement and twisting, as is evidenced by existence of special enzymes that reduce superhelicity of DNA and even unlock catenanes of circular DNA molecules by cutting and reconnecting the ends [4].

1.3. Get started

DNA, of course, is itself a double helix of two interwound sugar–phosphate chains completing one helical turn every 10.55 base pairs (bp) [5,6]. We are talking, thus, about folding of this duplex molecule. The DNA folding is assisted by proteins, in various degrees for different living kingdoms. In simple organisms, bacteria and viruses, DNA is, essentially, naked. In higher organisms, plants and animals, the folded state involves special proteins, in amount roughly equal to the DNA component. In these organisms the double-helical DNA winds locally into compact cylindrical (super)helices of only about 1.5 turns of the DNA, making remarkably standard tight packing units – nucleosome core particles, of 11 nm in diameter [7] (note that diameter of the DNA double helix is 2 nm). Their universal geometry is secured by a protein component, histone octamers, apparently nearly identical for the nucleosomes of all species, while DNA sections involved are all sequence-wise different. The particles are standard to such a degree that even their mixture, with different DNA sequences within the particles, can be crystallized [7].

The credit for discovery of the nucleosomes has to be given to four laboratories. First, Hewish and Burgoyne [8] discovered that enzymatic digestion of chromosomes results in fragmentation of DNA to about 200 bp long pieces and multiples of that size. This suggested that the chromosomes are built of units containing DNA of this size, and DNA connecting the units was somehow more accessible to the enzymes. The units themselves, in form of compact "ν-bodies", were then visualized by electron microscopy [9]. Physico-chemical disassembly–assembly studies by van Holde [10] and R. Kornberg [11] revealed that the particles consisted of four types of special proteins, histones, two molecules of each, and DNA wrapped around the *histone octamers*. The particles got the name "nucleosomes" [11].

One could think of random involvement of DNA in the nucleosomes, so that the sequence positions of the nucleosomes along any given gene would be uncertain and different for two identical copies of the gene. However, already in the 1970s it became clear that at least some nucleosomes do take specific positions along the sequences [12,13]. The latter work of Ponder and Crawford gave a strong clue as to what would be the sequence feature to which the histone octamers are attracted. It appeared that some nucleosomes could occupy several discrete positions on the same DNA fragment. Remarkably, these alternative positions were shifted one from another by 10–11 bp. Here the reader is suggested to exercise her (his) geometrical imagination. What the 10–11 bp shift would mean?

1.4. Degenerate sequence pattern versus unique pattern

The histone octamers, thus, seem to like special DNA fragments (sequences) with some common property. In other words, every such privileged fragment carries certain sequence signal expressed in some combination of letters (base pairs). Let us assume that the signal is unique and universal, sort of a label that appears in every 200 bp long

nucleosome DNA fragment. As simple combinatorics suggests, for the 4-letter alphabet of the nucleotide sequences (A, C, G and T) any 4-letter word, continuous or with fixed intervals between the letters, could serve as such label (occurrence of a given 4-letter word in fully random sequence is 1/256). However, no unique sequence signature of that kind has been ever detected in the nucleosome DNA sequences. The signal, thus, is far from being a unique sequence "word". One could imagine also some degenerate pattern, where certain letters would be preferable at certain positions (not obligatory, though). The number of such possible degenerate patterns is hopelessly high. Comparison of any two experimentally determined nucleosome sequences may show some similarity, but few other sequences added reduce the similarity (match) between all of them together to none. They all are seemingly as different as any set of unrelated, even random sequences would be. The sequence signal, or pattern, thus, got to be *highly* degenerate. That is, many completely different letter combinations may play the role of nucleosome positioning signal. It is like recognizing an animal in thick bushes: at each encounter one sees different parts of its image.

1.5. 10–11 base periodicity

We are now back to the effect of Ponder and Crawford. The discrete alternative positions of DNA on the surface of the histone octamer, every 10–11 bases, perhaps, indicate that these DNA fragments bind to the octamers *by the same side*. The DNA molecules, geometrically nearly identical to themselves with every base pair step shift along their axes and simultaneous rotation by 360°/10.55 around the axis – in accordance with their helical symmetry – should feel equally comfortable on the surface of the histone octamer after any number of such one base pair shifts and turns. Displacement by 10 or 11 steps, however, corresponds to full turn of the duplex around its axis, and thus alternative DNA fragments of Ponder and Crawford [13] find themselves facing the histone octamers by the same side. Somehow, the DNA molecule, thus, does have the side, the inner side, for certainty.

This sidedness can only be caused by some special distribution of bases in DNA. For example, placing certain bases, e.g., adenines (A) towards the interface between DNA and histones, that is, every 10–11 bases, may help specific interactions between the adenines and histones. No such base-to-histone links, however, have been detected so far.

Another simple thought is that since the base pairs of DNA are not strictly identical, nor the steps from one base pair to the next one, some of these steps may have their specific geometries slightly different from typical B-DNA average. For example, they may cause local deflection of DNA axis by a few degrees. If such axis deflecting steps are repeated every 10–11 bases, the DNA axis will be systematically turned in the same direction, thus, becoming curved. Such curved DNA would be, probably, good for winding it in the nucleosome.

Finally, one more possible property of DNA that may change with the period 10–11 bases is its deformability. Indeed, even if individual geometries of base-pair steps would be all identical (keeping DNA straight) some of the steps may be more deformable than others. Two base pairs making the step may become non-parallel due to easier opening, say, towards major groove of the DNA double helix, or towards phosphates. Placing such flexible element every 10–11 bases would make it also easier to bend DNA into an arc, making wider, e.g., all the major grooves where the deformable steps are located.

The notions of curved DNA and of sequence-dependent anisotropic deformability of DNA were first introduced in 1980 [14,15]. The respective sequence periodicity of DNA, that may reflect both the curvature and bendability, has been also first predicted and, indeed, detected in these works. Since then the DNA curvature and deformational anisotropy of DNA (two different notions! – see Section 2.6) have been considered as major sequence-dependent factors, probably, responsible for the nucleosome positioning along the DNA sequence. The 10–11 base periodicity of, primarily, AA and TT dinucleotides has been successfully used for sequence-directed mapping of the nucleosomes along the sequences, thus, introducing the new sequence code – chromatin code – that overlaps and coexists with the classical triplet code [15–18].

1.6. Helical period of DNA in the nucleosome is 10.4 bp per turn of the duplex

More than two decades since the discovery of approximately 10.5 base periodicity in eukaryotic (animals and plants) DNA sequences [14], the problem of exact value of the period has been a subject of debates. The helical repeat of the DNA duplex in the nucleosome is not necessarily the same as in free DNA. First, it may have a torsional

strain to store some energy to be used for unfolding of the nucleosome during DNA replication and transcription processes. Second, having an integer number of base pairs per turn in the nucleosome DNA (10 bp per turn, specifically) may render stabilizing interactions between adjacent superhelical turns of DNA in the nucleosome [19]. This value, 10 bp/turn, has been the first estimate for the structural period of the DNA in the nucleosome, dominating in literature since then. The third, and major, reason for the helical repeats of free DNA and nucleosome DNA being different is geometrical. The left-handed superhelical trajectory of the DNA axis in the nucleosome, as any curve in 3D space, is characterized at every point by curvature and torsion, both constant along an ideal superhelix. If Fuller–Crick formula [20] is applied to known geometrical parameters [7] of the DNA curve in the crystallized nucleosome, the torsion amounts to -0.15 bp/turn compared to the free DNA [21]. This is in good correspondence with the experimental measurement of average distance between nuclease digestion cuts in DNA within the nucleosome (10.35 ± 0.05 [22]) which is, indeed, less than helical repeat of free DNA, 10.55 ± 0.1 [5,6], by about the value above. The fit means that DNA, apparently, does not have much of additional energy-storing over/under-twisting in the nucleosome apart from the one dictated by the geometry of winding [21]. From DNaseI digestion experiments it is known that DNA in the nucleosome is cut once per every helical repeat [22], however, certain potential cut sites are rather skipped, as if the helical repeat of DNA would not be integer, and the cut sites would all have slightly different exposure, some being in an unfavorable orientation. From this "beat effect" the helical repeat of nucleosome DNA has been theoretically estimated in 1979 [23] to be 10.33–10.40 bp/turn. The final value for the nucleosome DNA helical period has been established only recently [24], by direct analysis of crystallographic coordinates of the sugar–phosphate chains of DNA in the nucleosome: 10.40 ± 0.04 bp/turn.

1.7. Chromatin folding unit

After over thirty years of chromatin studies the structure of elementary building block of chromatin, the tight complex of DNA with histones, approaches today to a certainty: 1.5 left-handed superhelical turns of 125 bp of DNA with helical repeat 10.4 bp/turn, wrapped around the histone octamer. It follows from several solved crystal structures of the core particles [25] and from analysis of phosphate-to-phosphate distances in the crystals [24]. Low-resolution neutron scattering from the crystals [26] revealed that the protein mass of the particles concentrates along a helical path parallel to the path of DNA, on its inner side. That turns the unit particle into a topologically linear structure with DNA and histones in side-by-side arrangement [27]. Its' partial or full unfolding, transition from the compact composite helical form to extended linear form is suggested by the above, perhaps, as one of structural stages in replication and/or transcription processes.

The elementary folding unit of chromatin, as described, should not be confused with the nucleosome, nor with the nucleosome core particle as it is presented in textbooks. Fig. 1 explains the difference. Note that the unit does not contain any free tails of DNA unbound to the histones. The unbound DNA rather belongs to the linkers connecting the elementary folding units of the chromatin.

1.8. Towards higher order structure

The linkers between neighboring units would define not only distance between them, but their relative orientation in 3D space as well. To make the linker longer by one base pair one has to rotate the next nucleosome by $360°/10.55 \sim 34°$ around the linker's axis. This became clear after experiments of Noll et al. [29] in which predominant linker lengths in the short linker range have been determined, about 7 and 17 bp (the middle positions of respective electrophoretic bands). The effect is explained by considering possibility that at some (sterically forbidden) linker lengths and, respectively, orientations of the neighboring nucleosomes, they would have to bodily penetrate into one another. The range 5–10 bp (5.27–10.55 in modern terms) with the average 7.5 (7.92) would be sterically permissive, as well as longer linkers, by increments of 10.55 bp: 18.47 and 29.02. All linkers longer than that would allow any orientations, as the nucleosomes are now sufficiently far away from one another.

Thus, building the higher order structure would involve the elementary folding units connected by sterically allowed linkers and respectively rotated around the linker DNA axes. The knowledge of exact positioning of the nucleosomes (hence, linker lengths) along the DNA sequence becomes a must, to make the 3D reconstruction of higher order structure possible.

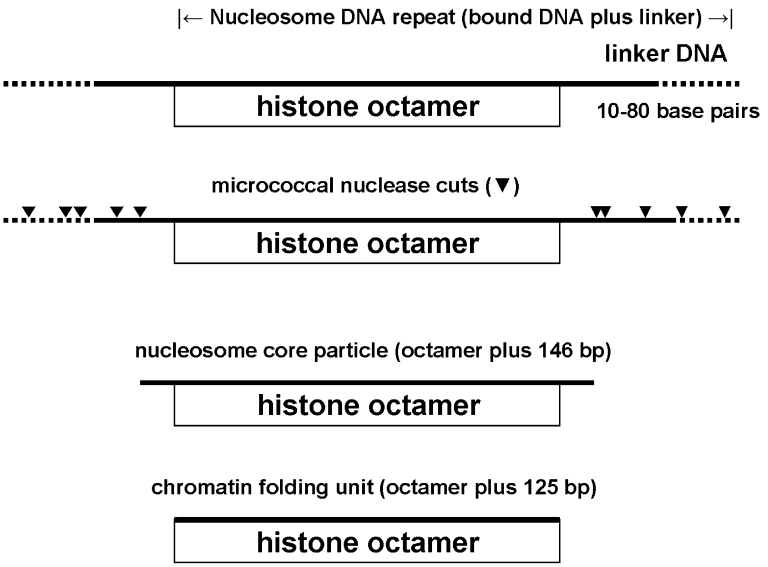


Fig. 1. Schematic presentation of the nucleosome DNA repeat, nucleosome core particle and elementary chromatin folding unit. Boxes – histone octamers. Thick horizontal lines represent DNA. The extra 21 bp of DNA in the core particles as compared to the units below remain undigested by micrococcal nuclease, apparently, because the very first base pairs of the linkers are sterically inaccessible to the nuclease.

1.9. Dinucleosome units

Dinucleosome units can be considered as next level, after the mononucleosomes, of the chromatin structure organization. They have been discovered as separate structural entities by a version of the same digestion technique that revealed the mononucleosomes, only using "bulky" nucleases, such that they could not reach and cut short DNA linkers [30,31]. Multiples of the dinucleosome size appeared in the gel, suggesting that there is an alternation of short and long linkers in the nucleosome arrays. These studies, unfortunately, did not receive proper recognition being buried under dominating views about 30 nm fibers of regular "solenoid" chromatin structure.

The traditional picture of the higher order structure of chromatin – solenoid model – has been speculated on the basis of low resolution data. Its regularity would imply that the nucleosomes are arranged along the sequences in periodical fashion. This expectation did not get experimental support. On the contrary, the nucleosomes appear to be placed at rather variable distances from one another [28]. Very recent high resolution electron microscopy study of chromatin structure in mitotic chromosomes conclusively demonstrated that the 30 nm fibers are not there [32]. The only characteristic size visible in the spectra of chromatin packaging, 10 nm, corresponds to the mononucleosomes.

Growing demand to clarify the undeniable role of higher order chromatin structure in gene regulation and expression, and recent insights in the nucleosome arrangements in the promoter regions of various organisms [33,34] urge new research efforts to elucidate the modes of folding and refolding of nucleosome arrays. Establishment of precise rules of nucleosome positioning along the sequences and, via linkers, in 3D space is, perhaps, the most promising approach to the higher order structure.

2. Deciphering the chromatin sequence code

2.1. AA/TT, but not only

The discovery of the AA/TT oscillating pattern, the first manifestation of existence of the chromatin code [15–18], has been confirmed in several subsequent studies, involving increasing sequence ensembles and whole genomes [35–38]. It came out, thus, a complete surprise, when human nucleosome DNA turned out practically not to contain the AA/TT oscillation [39,40] but rather RR/YY (R for A and G, Y for C and T) and CC/GG periodicity. The lonely earlier report on the CC/GG periodicity [41], thus, received support. First indications that these elements should be, perhaps, considered as well had been known, actually, since 1983 [17]. Yet another dimension has been opened by observation

that in genome of honey bee CG dinucleotides display very strong sequence periodicity of 10.4 bases, and the CG periodicity is the only periodical component visible in human genome [42]. This suggests that CG, as well as, perhaps, three other YR elements (CA, TA and TG) are additional nucleosome positioning sequence elements. That brings us back to the early idea of Zhurkin [43] that the YR and RY elements as dinucleotide steps of elevated bendability, especially towards the grooves of DNA duplex [44], would be important nucleosome positioning sequence steps. The suggestion had not been appreciated at that time, since no sequence periodicity of YR, nor RY steps has been detected, contrary to AA/TT and, later, GG/CC steps.

The full repertoire of the dinucleotide stacks that display the 10.4 base sequence periodicity and, thus, contribute to the nucleosome DNA bendability has been finally established only very recently [42]. It turned out that all 16 dinucleotides show the periodicity, although in various degrees, and not in all eukaryotic genomes, each showing its individual repertoire of the periodical dinucleotides (ibid).

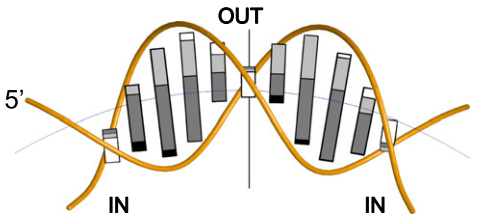
2.2. DNA bendability in the nucleosome

The sequence-dependent deformability of DNA is broadly accepted today as one of the major factors in the nucleosome positioning. This obviously physical property is not as simple as one would immediately imagine. It is not just bending rigidity of DNA as the molecular environment of DNA in the nucleosome is substantially different from conditions of free DNA in solution. A major difference is electrostatic asymmetry of the deformation. The phosphates of the inner side of the nucleosome DNA are neutralized by interactions with numerous positively charged amino acid residues of the histones. To measure or to theoretically estimate the deformability of the base pair stacks of the DNA within the nucleosome is not an easy task, especially for the stacks oriented orthogonally to the surface of the histone octamer, when both phosphates at the inner side are discharged.

One period long segment of the nucleosome DNA bent on the surface of the histone octamer contains 10 base (dinucleotide) positions. Each of the positions corresponds to a unique orientation of corresponding deformable base pair stacks relative to the surface of the histone octamer, differing by $\sim 34^\circ$ rotation with every base step. Each of 16 different base pair stacks (dinucleotides in the sequence) has its own deformational preferences, with easier opening either towards DNA grooves, or towards phosphates, or to some intermediate directions. Each of the stacks may be placed at its deformationally best position (orientation) within the nucleosome DNA helical period, thus helping the bending. A full description of the preferred positions for all 16 stacks within the period would boil down to the matrix with 16 lines (for the dinucleotides) and 10 columns (positions within the DNA period). This matrix has been called matrix of DNA bendability [17]. Placing the best dinucleotide at every position would result in high bendability of corresponding 10 base pair DNA segment. Smaller amount of the optimally oriented stacks within the period would also facilitate the bending but to a lesser degree. Thus, similarity of a given sequence to the matrix of bendability, from marginal to very high, would determine how well the sequence suits for the nucleosome formation. Note that if only a subset of the dinucleotides of the nucleosome DNA is optimally positioned, the preferential bending of DNA in certain direction may be secured in a given nucleosome by only one type of the dinucleotides, say, CG, periodically repeating along the DNA. One also could imagine a non-periodical sequence arrangement when every helical repeat of the nucleosome DNA would contain, say, one specific optimally positioned dinucleotide, but different for each repeat.

2.3. A little worm a rescuer

The efforts towards derivation of the complete nucleosome positioning pattern (matrix of bendability in our terms) formed by the dinucleotides have thirty years of agonizing history. Many tentative patterns have been suggested, very much different from one another (reviewed in [45]). All, naturally, have been met with skepticism, including our own suggestions, so that there is no broadly recognized sequence-directed routine for accurate detection of the nucleosome positions. (In this review biased towards our work, many significant contributions of other groups are not mentioned, largely for the sake of brevity. The reader may consider some of the omissions unfair, but these are not intentional.) Some aspects of the hidden signal, such as its sequence period and major contributing elements, could be outlined already on the basis of 200 [35] and 1300 [39] nucleosome DNA sequences. For the detailed picture, however, one had to wait for much larger ensemble. The frustratingly elusive nucleosome positioning sequence pattern could be fully extracted only when a very large database of the nucleosome DNA sequences became available, beating the




```
Y R R R R R Y Y Y Y Y R

S S     W W W W     S S

C G R A A A T T T Y C G

Y R       R Y       Y R

 IN       OUT       IN
```

Fig. 2. One helical repeat of the nucleosome DNA with the sequence YRRRRRYYYYY bent on the surface of the histone octamer (scheme). Purine bases are shown in dark grey. Binary forms of the chromatin code and its complete 4-letter alphabet version are shown aligned, with YR and RY dinucleotides centered at respective local dyads passing *between the bases* (note that since the helical period of the nucleosome DNA is not integer, the dyads may also pass *through bases* of YR and RY as shown in the scheme). Coordinates of sugar–phosphate chains and base pairs for the scheme are taken from [55].

overwhelming sequence noise. Such large collection, of about 160,000 sequences, has been produced in 2006 [46]. This was obtained by digestion of chromatin of small nematode *C. elegans*, and it soon became a conclusive source for establishment of the complete nucleosome DNA bendability matrix [47], first of the kind, which appears now to be universal for all eukaryotes (see below).

Contrary to earlier expectations according to which all 16 × 10 elements of the bendability matrix would comparably contribute to the DNA bending, the actual number of major contributing elements turned out to be about tenfold less. The matrix can be fairly accurately approximated by an elegant linear form CGRAAATTTYCG where complementarily symmetrical CG dinucleotides correspond to the minor grooves of DNA in contact with the surface, while other such symmetrical elements, AT dinucleotides, are 5 bases away (half-period of DNA), centered at the minor grooves oriented outwards [48,49] (see Fig. 2). The motif has perfect complementary symmetry which also means complete identity of two strands of the DNA duplex read in opposite directions.

The thirty years old history of the emerging chromatin code can be traced now in the following table:

xxAAAxxTTTxx	1980 [14,15], 1996 [35]
xxRRRxxYYYxx	1983 [17], 2003 [39]
YRxxxRYxxxYR	1983 [43]
YRRRRRYYYYYR	2008 [50]
xxxAAATTTxxx	2008 [50]
SSxxWWWWxxSS	2009 [51]
CGRAAATTTYCG	2009 [47]

The consecutive approximations converge with surprising loyalty to the final DNA bendability motif to which all the approximations fit. All these years the faint traces of the signal stirred imagination, often yielding to noise phantoms, and there was no way to derive the accurate pattern other than by, finally, exploring very large sequence databases. Or was there other way?

2.4. *A posteriori revelations*

When one scrutinizes the above code, naturally, the question arises: Why this unique pattern? Is it consistent with the idea of DNA bendability? Amazingly, the final pattern could be, actually, theoretically predicted provided

enough courage and imagination. With his share of shame the author has to admit the lack of the theoretical vision that prevented the right idea to appear already many years back. Here is the belated theoretical prediction, or rather *postdiction* [48]:

Stability (and rigidity) of DNA helix crucially depends on the intramolecular interactions in DNA, especially on stacking interactions between its base pairs [52]. When the duplex is deformed, the base pair stacks are deformed as well. Those of them that are positioned further away from the center of DNA bending suffer larger deformation. Bases and base pairs are moved apart from one another, working against the stacking interactions. The bases A and G (R, purines) in RR*YY stacks should be placed closer to the protein surface, as they are harder to deform (unstack) than interacting bases C and T (Y, pyrimidines) in complementary YY dinucleotides [53] (see Fig. 2). Similarly, the base pair stacks of the group WW*WW (AA*TT, TA*TA and AT*AT) of weak (W) bases A and T are generally less stable than the stacks SS*SS (GG*CC, CG*CG and GC*GC), with strong (S) bases C and G [54]. This means that the dinucleotides AA, TT, AT and TA should be put closer to exterior, that is at and close to the minor grooves of DNA oriented outwards (see Fig. 2), as they would require less energy for the unstacking. Similarly, the dinucleotides CC, GG, GC and CG should be put at the minor grooves oriented inwards. Thus, the ideal deformable nucleosome DNA sequences should follow the oscillating pattern ...SSSSSWWWWWSSSSSWWWWW.... Actually, as the eukaryotic genomes are A+T rich, the pattern would be rather ...SSSSWWWWWWSSSSWWWWWW.... That brings us to the following sequence arrangement required by minimization of the unstacking (Fig. 2):

```
5′-...SSSSWWWWWWSSSSWWWWWW...   and
5′-...YYRRRRRYYYYYRRRRRYYY...
```

The chromatin code, DNA bendability pattern, should satisfy both SS/WW and RR/YY binary patterns. The solution for this simple letter-choice problem for the alignment above is:

```
5′-...CCGGAAATTTCCGGAAATTT...   or
 5′-...CGGAAATTTCCG...   to compare with
       CGRAAATTTYCG
```

which is the pattern derived by massive DNA sequence analysis described in the previous section. Would the minimization of unstacking pattern be derived years back, all time- and effort-consuming computational work would have become unnecessary, except for its confirmatory value.

Interestingly, there is one more formally possible complementarily symmetrical solution for the binary patterns above [45]. It is obtained by their alternative alignment, with RY under SSSS and YR under WWWWWW (i.e. 5-base shift in the alignment above). The solution in this case is GCYTTTAAARGC, or TAAARGCYTTTA, inconsistent with the DNA deformability reasoning, that requires RY to reside within WWWW and YR – within SSSS (Fig. 2). Still, since the elements GC, CT, TA and AG do show the 10–11 nucleotide sequence periodicity in some genomes [42], the second symmetrical pattern may represent a less frequent alternative of the dominant nucleosome positioning sequence. Both, however, satisfy generalized pattern YRRRRRYYYYYR [17,50].

The complementary symmetry of the DNA deformability pattern CGRAAATTTYCG, and minimization of unstacking requires the RY (AT) dinucleotides to be positioned at the minor groove facing outwards (Fig. 2), while the dinucleotides YR (CG) are located in the minor groove oriented inwards. Inward position of YR dinucleotides is also confirmed by detailed calculations of the geometry of the YR*YR stacks within the nucleosome [56]. If the dinucleotides TA are used for the nucleosome DNA design, they have to take the same positions as CG, that is, inwards. This is exactly where these dinucleotides are found in the crystallized nucleosomes containing TA-periodical DNA [57].

The linear form of the DNA bendability pattern invites another surprisingly simple, “linguistic”, computation that only needs the values of observed frequencies of oligonucleotide words in the analyzed genome. Indeed, existence of the, apparently, universal hidden pattern spread all over eukaryotic genomes should exert certain pressure on the oligonucleotide composition of the genomes. Perhaps, some of, say, trinucleotides would appear in excess, being part of the imposing pattern. Indeed, in practically all eukaryotic genomes the trinucleotides AAA, AAT, ATT and TTT are the most frequent ones. Together they make, obviously, the strings AAATTT and AAAATTTT. In most of the genomes the triplets NAA are dominated by GAA, and triplets TTN – by TTC. This makes already 8 bases – GAAATTTC,

or even all 10 bases – GAAAATTTTC, of the complete $(GRAAATTTYC)_n$ periodical motif. This technique of pattern reconstruction (N-gram extension) has been first suggested by Shannon in 1948 [58]. It is currently in use for derivation of dominant sequence patterns for any sequence type (genome) of interest [59].

2.5. *Predictive potential of the nucleosome positioning pattern*

Displacement of the nucleosome DNA by 1–2 bases on the surface of the histone octamer would cause respective rotation of DNA around its axis by 34–68°. That would substantially change the direction of bending, away from the optimal one, suggested by the cumulative contribution of properly oriented base pair stacks along the molecule. In other words, sequence-directed predictions on the basis of the bendability pattern should be as sensitive, allowing, perhaps, just one–two bases uncertainty. In order to test this potential one has to rely upon experiments providing highly accurate nucleosome positions. Such essentially absolute accuracy is characteristic of the crystallized nucleosomes in which the central base of the nucleosome DNA sequence is determined with atomic resolution [25]. By repeating the derived DNA bendability pattern, with correction for the non-integer 10.4 base sequence period, the nucleosome sequence "probe" has been designed for the nucleosome mapping purposes [49,60]. Only seven cases are available for the accuracy testing purposes [49]. They all demonstrated either none or just one base misfit with the calculated positions. Together with the fact that the same bendability pattern is now derived by three completely different approaches [47,48,59], this comparison gives a good measure of confidence in the new sequence-directed nucleosome mapping technique. It can only be challenged by similar testing of other known sequence-based mapping techniques, e.g. [35,61], on the highly accurately mapped nucleosomes. To our knowledge, no such testing have been made so far. Rather, the computational predictions have been compared with experimental maps obtained by micrococcal nuclease digestion, that is with uncertainty of 10–20 bases [46]. The *correlation* has been observed, indeed [61], but the *fit* could not be checked on the basis of the low resolution data.

2.6. *Is DNA curvature involved in the nucleosome positioning?*

Discussion of this difficult question is reserved for the very end of the review, to minimize an inevitable confusion, since some aspects of the involvement of the DNA curvature in the nucleosome formation are rather counterintuitive. Quite often the question of the subtitle is met with a surprise: "But is not it the same – DNA bending and DNA curvature?" It is not.

DNA curvature is property of free unconstrained DNA. The DNA axis in this case is curvilinear because of accumulation of small differences in the angles between neighboring essentially flat base pairs, caused by differences in their chemical structure [62,63]. DNA bending, on the other hand, is *deformation* of the DNA axis due, e.g., to binding of DNA to some proteins, in particular, wrapping around the histone octamers. Both curvature and bending are sequence dependent, but in different ways, as static geometry of DNA and its deformation are, indeed, different things.

Curved DNA may form rather stable nucleosomes [64,65]. Incidentally, the DNA with repeating ideal nucleosome positioning sequence GAAAATTTTC, as above, does possess an intrinsic curvature. This is experimentally observed, by characteristic anomalous electrophoretic mobility of this DNA [66,67]. The direction of the curvature is such that the AATT tetranucleotide is located in minor groove oriented inwards, facing the center of the arc. This follows both from the fact that AA and TT dinucleotides cause narrowing the minor groove of DNA [68], and from the calculations of the dinucleotide wedge components on the basis of large amount of DNA electrophoretic mobility data [67,63]. However, the bending direction of this very DNA *in the nucleosome* is (sic!) exactly opposite, with central AATT oriented outwards, away from the center of the arc of the deformed DNA. This follows from all the data outlined in the sections above, especially from [48,56,49]. The self-complementarity, i.e., dyad symmetry of the duplex with the sequence GAAAATTTTC demands that both vectors of the curvature and of preferred direction of deformation should be collinear with the local dyad at the AT step, with either the same or opposite sign. The latter is the case. The inverse direction of the curvature in the GAAAATTTTC DNA duplex not bound to the histones suggests that, perhaps, such sequence arrangement facilitates "peeling" the DNA off the histone octamer during transcription and replication.

The DNA bendability motif GAAAATTTTC has only poor similarity to the (non-symmetrical) motifs AAAAAACGCG [69] and AAAAATGACT [70] associated with the highest DNA curvature observed so far, further illustrating the dif-

ference between the DNA curvature and DNA bendability sequence patterns. The contribution of the DNA curvature to the nucleosome formation and unfolding is, thus, matter of future detailed studies.

2.7. To conclude

The periodical pattern of dinucleotides responsible for the DNA bending in the nucleosomes – chromatin code – appears to be, finally, established. It is derived by three independent approaches: by signal processing analysis of very large nucleosome DNA sequence database, by theoretical reconstruction via minimization of unstacking, and by Shannon N-gram extension.

The pattern is expressed in simple linear forms:

$(\mathtt{GRAAATTTYC})_n$ and
$(\mathtt{RRRRRYYYYY})_n$

which have two axes of complementary symmetry, at CG (YR) and at AT (RY) steps, separated by half-turn of DNA duplex.

2.8. Frequently asked questions

Q: Why one does not see that sequence in the nucleosome DNA?
A: One would see that pattern only in very strong nucleosomes. These, however, are very much avoided. In reality there are only of the order 10 signal dinucleotides per nucleosome from the magic repertoire, any of CC, CG, GG, GA, AA, AT, TT and TC scattered here and there along the nucleosome DNA, but in positions matching the periodical pattern. This provides some similarity to the repeated matrix of bendability and its simplified linear form, sufficient for formation of from moderately to marginally stable nucleosomes.

Q: Experiments are well known [71,72] which demonstrate that exceptionally stable nucleosomes can be made on sequences showing strong periodicity of TA dinucleotides. Why TA is not among the magic ones?
A: TA*TA stacks are least stable of all. When placed every 10 or so base pairs along the sequence they would form, indeed, strong nucleosomes, but with kinks [73,57] at TA steps. The kinks are mutational hot spots and should be avoided. Indeed, TA is second most avoided dinucleotide in eukaryotic DNA, and it displays 10–11 nucleotide sequence periodicity only very rarely [42].

Q: CG is the most avoided dinucleotide. Why it is, nevertheless, among the magic ones?
A: It is also most mutable (methylation of C) and avoided because of that. But it is very good for gene regulation by CG methylation and it is good for nucleosome formation. Apparently, opposite tendencies are balanced.

Q: What are the forces that cause bending of DNA in the nucleosome?
A: Since the contacts between the negative phosphates in the minor grooves and positive arginine residues on the surface of the histone octamer are electrostatic, the bending force, apparently, is electrostatic as well [74].

Q: Do not some other detailed sequence patterns suggested in literature represent legitimate alternative patterns?
A: The *linear* form of the code is a simplification. Full description of the code would involve other dinucleotide elements as well, with their preferred positions within the *matrix* of bendability [17,47]. Their combinations may appear as different pattern. However, every such pattern would still follow the same physical rules of deformability being just a partial description of the matrix.

Q: What do you think about the genomic code for chromatin of Segal et al. [61]?
A: The pattern suggested in [61], AA, TA, TT/GC, with dominant TA, resembles the possible minor alternative motif described above. However, marginal periodicity of TA in natural sequences [42] suggests that this pattern is of a limited use in the sequences.

Acknowledgements

The work has been supported by the grant 222/09 of Israel Science Foundation and by Fellowship of SoMoPro (South Moravian Program, Czech Republic) with financial contribution of European Union within the 7th framework program (FP/2007-2013, grant agreement No. 229603). Special thanks to Jan Hapala (Masaryk University) and Zakharia Frenkel (Haifa) for help in preparing figures.

References

[1] Borochov N, Eisenberg H, Kam Z. Dependence of DNA conformation on the concentration of salt. Biopolymers 1981;20:231–5.
[2] Suzuki M, Wakabayashi T. Packaging of DNA in cricket sperm: A compact mode of DNA packaging. J Mol Biol 1988;204:653–61.
[3] Jardine PJ, Anderson DL. DNA packaging in double-stranded DNA phages. In: Calendar R, editor. The bacteriophages. 2nd ed. Oxford: Oxford University Press; 2006. p. 49–65.
[4] Wang JC. DNA topoisomerases: why so many? J Biol Chem 1991;266:6659–62.
[5] Peck LJ, Wang JC. Sequence dependence of the helical repeat of DNA in solution. Nature 1981;292:375–8.
[6] Strauss F, Gaillard C, Prunell A. Helical periodicity of DNA, Poly(dA).poly(dT) and poly(dA-dT).poly(dA-dT) in solution. Eur J Biochem 1981;118:215–22.
[7] Richmond TJ, Finch JT, Rushton B, Rhodes D, Klug A. Structure of the nucleosome core particle at 7A resolution. Nature 1984;311:532–7.
[8] Hewish DR, Burgoyne LA. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. Biochem Biophys Res Commun 1973;52:504–10.
[9] Olins AL, Olins DE. Spheroid chromatin units (v bodies). Science 1974;183:330–2.
[10] Van Holde KE, Sahasrabuddhe CG, Shaw BR. A model for particulate structure in chromatin. Nucleic Acids Res 1974;1:1579–86.
[11] Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science 1974;184:868–71.
[12] Chao MV, Gralla JD, Martinson HG. DNA sequence directs the placement of histone cores on restriction fragments. Biochemistry 1979;18:1068–74.
[13] Ponder BAJ, Crawford LV. The arrangement of nucleosomes in nucleoprotein complexes from polyoma virus and SV40. Cell 1977;11:35–49.
[14] Trifonov EN, Sussman JL. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA 1980;77:3816–20.
[15] Trifonov EN. Sequence-dependent deformational anisotropy of chromatin DNA. Nucleic Acids Res 1980;8:4041–53.
[16] Trifonov EN. Structure of DNA in chromatin. In: Schweiger H, editor. International cell biology 1980–1981. Berlin: Springer-Verlag; 1981. p. 128–38.
[17] Mengeritsky G, Trifonov EN. Nucleotide sequence-directed mapping of the nucleosomes. Nucleic Acids Res 1983;11:3833–51.
[18] Trifonov EN. Sequence codes. In: Creighton TE, editor. Encyclopedia of molecular biology. New York: John Wiley & Sons, Inc.; 1999. p. 2324–6.
[19] Finch JT, Lutter LC, Rhodes D, Brown RS, Rushton B, Levitt M, et al. Structure of nucleosome core particles of chromatin. Nature 1977;269:29–36.
[20] Crick FHC. Linking numbers and nucleosomes. Proc Natl Acad Sci USA 1976;73:2639–43.
[21] Ulanovsky LE, Trifonov EN. Superhelicity of nucleosomal DNA changes its double-helical repeat. Cell Biophys 1983;5:281–3.
[22] Prunell A, Kornberg R, Lutter L, Klug A, Levitt M, Crick F. Periodicity of deoxyribonuclease I digestion of chromatin. Science 1979;204:855–8.
[23] Trifonov EN, Bettecken T. Noninteger pitch and nuclease sensitivity of chromatin DNA. Biochemistry 1979;18:454–6.
[24] Cohanim AB, Kashi Y, Trifonov EN. Three sequence rules for chromatin. J Biomol Struct Dyn 2006;23:559–66.
[25] Richmond TJ, Devey CA. The structure of DNA in the nucleosome core. Nature 2003;423:145–50.
[26] Bentley GA, Finch JT, Lewit-Bentley A. Neutron diffraction studies on crystals of nucleosome cores using contrast variation. J Mol Biol 1981;145:771–84.
[27] Trifonov E. The helical model of the nucleosome core. Nucleic Acids Res 1978;5:1371–80.
[28] Drew HR, Calladine CR. Sequence-specific positioning of core histones on an 860 bp DNA – experiment and theory. J Mol Biol 1987;195:143–73.
[29] Noll M, Zimmer S, Engel A, Dubochet J. Self-assembly of single and closely spaced nucleosome core particles. Nucleic Acids Res 1980;8:21–42.
[30] Burgoyne LA, Skinner JD. Chromatin superstructure: The next level of structure above the nucleosome has an alternating character. A two nucleosome based series is generated by probes armed with DNAase-I acting on isolated nuclei. Biochem Biophys Res Commun 1981;99:893–9.
[31] Khachatrian AT, Pospelov VA, Svetlikova SB, Vorobiev VI. Nucleodisome – a new repeat unit of chromatin revealed in nuclei of pigeon erythrocytes by DNase I digestion. FEBS Lett 1981;128:90–2.
[32] Maeshima K, Hihara S, Eltsov M. Chromatin structure: does the 30-nm fibre exist in vivo? Curr Opin Cell Biol 2010;22:291–7.
[33] Ioshikhes I, Trifonov EN, Zhang MQ. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc Natl Acad Sci USA 1999;96:2891–5.
[34] Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 2009;10:161–72.
[35] Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol 1996;262:129–39.

[36] Cohanim AB, Kashi Y, Trifonov EN. Yeast nucleosome DNA pattern: deconvolution from genome sequences of *S. cerevisiae*. J Biomol Struct Dyn 2005;22:687–94.
[37] Herzel H, Weiss O, Trifonov EN. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 1999;15:187–93.
[38] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18:1051–63.
[39] Kato M, Onishi Y, Wada-Kiyama Y, Abe T, Ikemura T, Kogan S, et al. Dinucleosome DNA of human K562 cells: experimental and computational characterizations. J Mol Biol 2003;332:111–25.
[40] Kogan SB, Kato M, Kiyama R, Trifonov EN. Sequence structure of human nucleosome DNA. J Biomol Struct Dyn 2006;24:43–8.
[41] Bolshoy A. CC dinucleotides contribute to the bending of DNA in chromatin. Nat Struct Biol 1995;2:446–8.
[42] Bettecken T, Trifonov EN. Repertoires of the nucleosome-positioning dinucleotides. PLoS One 2009;4:e7654.
[43] Zhurkin VB. Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine–pyrimidine and pyrimidine–purine dimers. FEBS Lett 1983;158:293–7.
[44] Zhurkin VB, Lysov YP, Ivanov VI. Anisotropic flexibility of DNA and the nucleosomal structure. Nucleic Acids Res 1979;6:1081–96.
[45] Trifonov EN. Nucleosome positioning by sequence, state of the art and apparent finale. J Biomol Struct Dyn 2010;27:741–6.
[46] Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. Genome Res 2006;16:1505–16.
[47] Gabdank I, Barash D, Trifonov EN. Nucleosome DNA bendability matrix (*C. elegans*). J Biomol Struct Dyn 2009;26:403–12.
[48] Trifonov EN. Base pair stacking in nucleosome DNA and bendability sequence pattern. J Theor Biol 2010;263:337–9.
[49] Gabdank I, Barash D, Trifonov EN. Single-base resolution nucleosome mapping on DNA sequences. J Biomol Struct Dyn 2010;28:107–21.
[50] Salih F, Salih B, Trifonov EN. Sequence structure of hidden 10.4-base repeat in the nucleosomes of *C. elegans*. J Biomol Struct Dyn 2008;26:273–82.
[51] Chung HR, Vingron M. Sequence-dependent nucleosome positioning. J Mol Biol 2009;386:1411–22.
[52] Anselmi C, Bocchinfuso G, De Santis P, Savino M, Scipioni A. A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. Biophys J 2000;79:601–13.
[53] Svozil D, Hobza P, Sponer J. Comparison of intrinsic stacking energies of ten unique dinucleotide steps in A-RNA and B-DNA duplexes. Can we determine correct order of stability by quantum-chemical calculations? J Phys Chem B 2010;114:1191–203.
[54] Krueger A, Protozanova E, Frank-Kamenetskii MD. Sequence-dependent base pair opening in DNA double helix. Biophys J 2006;90:3091–9.
[55] Sussman JL, Trifonov EN. Possibility of non-kinked packing of DNA in chromatin. Proc Natl Acad Sci USA 1978;75:103–7.
[56] Wang D, Ulyanov NB, Zhurkin VB. Sequence-dependent kink-and-slide deformations of nucleosomal DNA facilitated by histone arginines bound in the minor groove. J Biomol Struct Dyn 2010;27:843–59.
[57] Vasudevan D, Chua EY, Davey CA. Crystal structures of nucleosome core particles containing the '601' strong positioning sequence. J Mol Biol 2010;403:1–10.
[58] Shannon CE. A mathematical theory of communication. The Bell System Technical Journal 1948;27:379–423; 623–56.
[59] Rapoport AE, Frenkel ZM, Trifonov EN. Nucleosome positioning pattern derived from oligonucleotide compositions of genomic sequences. J Biomol Struct Dyn 2011;28:567–74.
[60] Gabdank I, Barash D, Trifonov EN. FineSTR: a web server for single-base-resolution nucleosome positioning. Bioinformatics 2010;26:845–6.
[61] Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature 2006;442:772–8.
[62] Trifonov EN. Curved DNA. CRC Crit Rev. Biochem 1985;19:89–106.
[63] Bolshoy A, McNamara P, Harrington RE, Trifonov EN. Curved DNA without AA: experimental estimation of all 16 wedge angles. Proc Natl Acad Sci USA 1991;88:2312–6.
[64] Pennings S, Muyldermans S, Meersseman G, Wyns L. Formation, stability and core histone positioning of nucleosomes reassembled on bent and other nucleosome-derived DNA. J Mol Biol 1989;207:183–92.
[65] Shrader TE, Crothers DM. Effects of DNA sequence and histone–histone interactions on nucleosome placement. J Mol Biol 1990;216:69–84.
[66] Hagerman PJ. Sequence-directed curvature of DNA. Nature 1986;321:449–50.
[67] Ulanovsky LE, Trifonov EN. Estimation of wedge components in curved DNA. Nature 1987;326:720–2.
[68] Nelson HCM, Finch JT, Luisi BF, Klug A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature 1987;330:221–6.
[69] Koo HS, Crothers DM. Calibration of DNA curvature and a unified description of sequence-directed bending. Proc Natl Acad Sci USA 1988;85:1763–7.
[70] Haran TE, Kahn JD, Crothers DM. Sequence elements responsible for DNA curvature. J Mol Biol 1994;244:135–43.
[71] Widlund HR, Cao H, Simonsson S, Magnusson E, Simonsson T, Nielsen PE, et al. Identification and characterization of genomic nucleosome-positioning sequences. J Mol Biol 1997;267:807–17.
[72] Lowary PT, Widom J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol 1998;276:19–42.
[73] McNamara PT, Bolshoy A, Trifonov EN, Harrington RE. Sequence-dependent kinks induced in curved DNA. J Biomol Struct Dyn 1990;8:529–38.
[74] Mirzabekov AD, Rich A. Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending. Proc Natl Acad Sci USA 1979;76:1118–21.