An algorithm for studying cooperative transitions in DNA

H.Marcaud, J.Gabarro-Arpa, R.Ehrlich and C.Reiss

Institut Jacques Monod, CNRS and Université de Paris VII, tour 43, 2 place Jussieu, 75251 Paris, France

Received 22 July 1985

SUMMARY: Cooperative transitions in DNA (B to Z, B to A, helix to coil, etc.) are known to depend strongly on nucleotide sequence. In general the change in free energy involved in the transition can be expressed as:

$$\Delta G_{(seq)} = 2RT \log(\sigma)$$

where σ is a factor arising from the free energy associated with boundaries of different conformations along the molecule.

This formula allows to infer a general algorithm with which DNA sequences can be partitioned into well defined domains in which, under suitable conditions, base pairs change state cooperatively.

The different partitions of the sequence that can be generated by varying the values of the physical parameters involved in the above formula, are shown to be embedded into a binary tree hierarchy.

Application to a reliable prediction of Z-DNA antibody binding sites will be illustrated for the ϕ X174 genome. Possible biological implications are briefly discussed. (*)

INTRODUCTION

Nucleic acids may adopt a number of conformations: A, B, Z, coil, etc. and under appropriate physical conditions they can readily switch between two of these conformations: helix to coil, B to Z, B to A, etc. (1-5). These conformations, and transitions between two of them, are highly dependent on base composition since the value of the parameters monitoring the transition point (supercoiling, hydration, temperature etc.) varies over large ranges for different polymers (5-7).

In natural DNA, alternating conformations were observed to appear in domains i.e. continuous base pair (bp.) stretches with well defined location inside the sequence sharing a given conformation. Domain sequences allow for "mismatches": disrupted GC bp. in an AT rich coiled segment below the poly(GC) melting point (3), Z segments in supercoiled DNA with breaks in the alternance

(*) Programs mentioned in this paper are available on request.

pur-pyr (8), etc.

This phenomenon may be explained by the fact that for each boundary between two conformations (helix-coil, B-Z, A-B, ...) there is an energy term which adds a sizable contribution to the difference in free energy between the two conformations (2,4,5,10); thus the number of boundaries is kept to a minimum.

The basic idea behind the algorithm developed below is that by adding up the free energies of the different elements in the sequence, the stretches that have a total free energy negative enough to match the boundary energy build domains, provided they are neither embedded in nor contain segments of higher positive free energy. We emphasize that this procedure is basically different from the various usual averaging or smoothing procedures.

Four parts follow this introduction:

In the first part, there is a detailed description of the algorithm (originally proposed by Azbel (11)) showing how a sequence can be partitioned into domains.

Follows a second part where we show how the set of partitions that can be generated is hierarchically organized into a binary tree.

In the third part, it is shown how this procedure can be used to predict the occurrence of anti-Z DNA immunoglobulin G (IgG) binding sites in supercoiled DNA; it has also been successfully applied to the description of the helix-coil transition occurring in thermal DNA melting (1-3, and refs. therein) and to the finding of homologous domains between unrelated sequences (12,13).

Lastly, follows a discussion on developments and possible biological role of domains.

DESCRIPTION OF THE ALGORITHM

The description we give here aims to be simple and general. For each particular application, this algorithm can be modified and adapted.

The sequence we are going to deal with is built of two kinds of elements: 0s and 1s, they can stand for A/T and G/C bp. in helix coil transition; alternance pur-pyr and non alternance in B-Z transition, mismatch and match in sequence comparison etc...

Let X be some parameter monitoring the transition between two states: a to β , and let the differences in free energy between them be:

$$\Delta F_0 = \Delta S \cdot (X - X_0) \quad \Delta F_1 = \Delta S \cdot (X - X_1) \quad -1-$$

where ΔS is the change in entropy (we take ΔS to be the same for both a and

β), and X_0 (resp. X_1) the transition point for a poly(0)(resp. poly(1)) sequence. All transitions take place between $X_1 \rightarrow X \rightarrow X_0$, with $\Delta F_0 > 0 > \Delta F_1$.

Let us introduce the dimensionless parameter $p = (X - X_0) / \Delta X$ ($\Delta X = X_1 - X_0$), the free energy of an α stretch of length L can be written:

$$E = \Delta S \cdot \Delta X \cdot [n_0 p + n_1 (p-1)] = \Delta S \cdot \Delta X \cdot L(p-r) = \Delta S \cdot \Delta X \cdot H \quad -2-$$

where n_0 (resp. n_1) is the number of elements of 0s (resp. 1s), $L = n_0 + n_1$, $r = n_1 / L$, H is a renormalized dimensionless variable which, for the sake of simplicity, H will be used instead of E . The energy of a stretch in β state is given the zero value. Moreover, for each boundary between segments in different states we should add a term $W = 2V / (\Delta S \cdot \Delta X)$ (V being the boundary free energy).

For each given sequence, let us build the following recursive function:

$$h(0) = 0 \quad ; \quad h(m) = h(m-1) + \begin{cases} p & \text{if the } m^{\text{th}} \text{ is } 0 \\ p-1 & \text{" " " " " 1} \end{cases}$$

so for segment (m, n) :

$$H(m, n) = h(n) - h(m) \quad -3-$$

Let us consider the following hypothetic sequence:

$$(1)_1^1 (0)_{30}^2 (1)_{60}^3 (0)_{40}^4 (1)_{42}^5 (0)_{6}^6 (1)_1^1, \quad 1, 1' > 120$$

composed of seven alternate segments with multiplicity indicated as a subscript (in what follows numbers 1 to 6 will indicate boundaries instead of element positions). The h function for this sequence is depicted in fig. 1a. Also assume for the present example that $W = W_1$ ($=15$, see fig. 1a). Three kinds of transitions arise:

a) for segment [1-2] we have:

$$H(1, 2) = h(2) - h(1) \quad \begin{cases} < \\ = W_1 & \text{for } p \\ > \end{cases} \quad \begin{cases} < \\ = .5 \\ > \end{cases}$$

in order to minimize its free energy, this segment will be in α state for $p < .5$ and in β for $p > .5$. The formation of this β segment in an α environment creates two α - β boundaries.

b) for segment [2-3] the total H value will be:

$$H(2, 3) = h(3) - h(2) + W_1 \quad \begin{cases} < \\ = 0 & \text{for } p \\ > \end{cases} \quad \begin{cases} < \\ = .75 \\ > \end{cases}$$

the term W_1 is added here because, for $p > .5$, neighbour segments are in β state adding thus two boundaries (see fig. 1b). Again to minimize its free energy,

Figure 1.

a) h function for the sequence used as an example, depicted for $p = .375, .5, .666, .75, .875$. Absolute vertical position is irrelevant as only height differences matter. Numbers in abscissae refer to boundaries between 0-1 stretches. $W_1=15$, $W_2=5.25$, $W_3=53.75$.

b) successive conformations in the sequence as p is continuously raised, with $W=W_1$. Solid lines represent segments in α state, dotted lines segments in β state. ∇ indicate newly created boundaries, $\blacktriangledown$ stands for destroyed boundaries.

for $p > .75$, this segment will be in β state. In this kind of transition two boundaries disappear.

c) for segment [4-6] we have a transition for $p > .875$:

$$H(4,6)=h(6)-h(4) \begin{cases} < \\ = 0 \text{ for } p \\ > \end{cases} \begin{cases} < \\ = .875 \\ > \end{cases}$$

as segment [1-4] is already in β state for $p > .875$, [4-6] in α state will add positive free energy to the system while adding zero in β state. This transition leaves the number of boundaries unchanged.

All three kinds of transitions can be characterized by the formula

$$L(p-r)=nW \quad -4-$$

where n is a boundary coefficient: a) $n=+1$, b) $n=-1$, c) $n=0$.

We have taken here a fixed value for W , which suffices in many

Figure 2.

a) W vs. p plot. Solid lines are transitions represented according to -4-, ● indicate branching points. This plot is actually a phase diagram since lines delimit areas where the polymer is in a well defined minimal energy state.

b) The same plot with sequence added as a third dimension.

applications. However V is known to depend on the environment (2) so we are going to see below what happens when W varies over a continuous positive range.

HIERARCHICAL ORGANIZATION OF SEQUENCE PARTITIONS

If we have $W > W_3 (=53.75)$, see fig. 1a), the set of transitions depicted in fig. 1b are energetically unfavourable and can no longer be obtained. Only segment [1-6] can undergo a transition as $H(1,6)=W$ for $p > .875$, some subsegments have a negative H , but as their magnitude is less than that of the total segment, the transition of the segment as a whole is energetically favourable.

Now for $W=W_3$, $p=.875$ (see fig. 1a) segment [1-6] can undergo a transition either in one step or in two steps: first an a) transition for segment [1-4] followed by a c) transition in segment [4-6]. For $W < W_3$, the latter scheme will prevail. So by decreasing W , an a) transition can split into an a) and a c) transition.

For $W=W_2 (=5.25)$ in the example of fig. 1a, segment [4-6] can change state by following either of two paths: by a c) transition in one step, or in two steps an a) transition for segment [5-6] followed by a b) transition for segment [4-5]. Again for $W < W_2$ the latter two step process prevails.

As a rule a transition will split everytime W becomes small enough to allow a subsegment to have the same H magnitude as the whole segment. Parent

Figure 3.

Map of ϕ X174: bases are numbered clockwise, convention used is the same as in (17), major restriction sites are indicated; black areas cover the 9 anti-Z IgG binding sites identified in (9); external arrows indicate the predicted sites see table below.

transitions having boundary number n will split into two daughter transitions with boundary numbers such that $n = n_1 + n_2$ (2). When W is allowed to take any positive value, the set of transitions generated from a given sequence is organized into a binary tree hierarchy, this can be seen in a plot of W vs. p , see fig. 2a for the present example, if the sequence is added as a third dimension this plot appears as a solid surface in fig. 2b.

DETERMINATION OF ANTI-Z IgG BINDING SITES

Double stranded DNA in B form is characterized by *anti* conformation of the bases relative to the sugar, whereas in Z form, every other base adopts the *syn* conformation. For purines, the energy difference between *syn* and *anti* was evaluated to be approximately 2 kcal mol^{-1} , but for pyrimidines, it amounts to $5\text{--}7 \text{ kcal mol}^{-1}$ (16). Thus, in contrast to pyrimidines, purines can easily switch from *anti* to *syn* conformation, so that Z forms are found primarily among alternating pur-pyr stretches; *syn* conformation of pyr is only expected if some other stabilization factors favour it (15).

We applied the above algorithm in assigning to pur-pur or pyr-pyr stacks ("1") a free energy differing from that of pur-pyr or pyr-pur stacks ("0") by $3.5 \text{ kcal mol}^{-1}$. Furthermore, the V value of the B-Z interface was taken, from (4), to be 5 kcal mol^{-1} , hence $W \sim 2.5$. Domains in the "conformability" diagram (p_c vs. seq.) range from highly non alternating (high p_c) to highly

TABLE I.

Z-sites from Electron Micr.	Z-sites from computation	p^s	p^c
180±85	248-251	.625	.29
988±49	955-959	.57	.33
	977-994	.60	.33
1422±47	1374-1387	.57	.42
	1391-1399	.75	.42
	1430-1436	.67	.25
2230±42	2185-2197	.69	.42
	2232-2245	.57 to .63	.15
2496±20	2496-2506	.62 to .63	.23
3223±20	3287-3293	.67	.33
3581±100	3531-3535	.60	.31
	3632-3634	.57	.38
3832±31	3832-3837	.62 to .67	.42
	4263-4274	.71 to .75	.33 to .37
5013±62	5020-5032	.58	.33 to .43

alternating (low p_c), the latter being candidates to Z form.

It has however been found experimentally that the pur-pyr alternance is not sufficient to enable the Z form (7); the GC content in the pur-pyr alternation must predominate, presumably because the *syn-anti* transition of adenines is less favorable than that of guanines (14). Therefore, in practice we superpose a conventional stability profile (p_s vs. seq., (3)) to the conformability profile, and choose empirically minimum p_c (high alternation) and maximum p_s (high GC content). This procedure works well, but lack of enough experimental data prevents formalization of the above rule, which for the moment remains empirical.

As an example we apply these considerations to ØX174 so as to select the 9 Z sites identified by electron microscopy (9). For $p_c < .42$ (roughly over 58% alternation) and $p_s > .58$ (roughly over 58% GC), 10 domains are found, including the 9 determined experimentally (figure 3 and table I).

CONCLUSION

A body of evidence has accumulated in favor of domains in DNA so: do they play a role in regulating gene expression? is an obvious question many laboratories (including our own) are attempting to answer.

We have found (18,19) that characteristic genetic signals borne by nucleic acids, and controlling or regulating gene expression (promoters in particular), coincide with characteristic stability and/or conformability domains. These result from particular local constraints in the distribution of

AT/GC bp. or pur/pyr stacks; both constraints are independent and may be encoded by different synonymous sequences. It seems to us that these domains, which are under environmental control, need not to be permanent features of a particular DNA molecule but may arise as transient structures, as presumably is the case with Z DNA stretches, that can be recognized by target proteins.

The algorithm presented above exists in different versions contained in the programs FASE, PASE, DUR and PRINFAS, all written in MicroSoftTM BASIC.

ACKNOWLEDGEMENTS

We wish to thank F. Bontemps and B. Dalton for critical reading of the manuscript. One of us (H.M.) wishes to thank D.G.R.S.T. and Centre National de la Recherche Scientifique for financial support.

REFERENCES

1. Gabarro-Arpa, J., Tougard, P., Reiss, C., (1979) *Nature* **280**, 515-517.
2. Gabarro-Arpa, J., Michel, M., (1982) *Biochimie* **64**, 99-112.
3. Michel, F., Gabarro-Arpa, J., Dujon, B., (1982) *Biochimie* **64**, 113-126.
4. Peck, L.J., Wang, J.C., (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6206-6210.
5. Ivanov, V.I., Minchenkova, L.E., Minyat, E.E., Schyolkina, A.K., (1983) *Cold Spring Harbor Symposia on Quantitative Biology* vol. XLVII part 2, 243-250.
6. Marmur, J., Doty, P., (1962) *J. Mol. Biol.* **5**, 109-115.
7. Wang, A.J., Hakoshima, T., van der Marel, G., van Boom, J.H., Rich, A., (1984) *Cell* **37**, 321-331.
8. Nordheim, A., Lafer, E.M., Peck, L.J., Wang, J.C., Stollar, B.D., Rich, A., (1982) *Cell* **31**, 309-318.
9. Revet, B., Zarling, D.A., Jovin, T.M., Delain, E., (1984) *EMBO J.* **3**, 3353-3358.
10. Frank-Kamenetskii, M.D., Chogovadze, G.I., (1984) *J. Biomol. Struc. & Dyn.* **1**, 1517-1523.
11. Azbel, M.Ya., (1973) *Phys. Rev. Lett.* **31**, 589-592.
12. Michel, F., Jacquier, A., Dujon, B., (1982) *Biochimie* **64**, 867-881.
13. Michel, F., Dujon, B., (1983) *EMBO J.* **2**, 33-38.
14. Jovin, T.M., McIntosh, L.P., Arndt-Jovin, D.J., Zarling, D.A., Robert-Nicoud, M., van de Sande, J.H., Jorgenson, K.F., Eckstein, F., (1983) *J. Biomol. Struc. & Dyn.* **1**, 21-57.
15. Wang, A.J., Gessner, R.V., van der Marel, G.A., van Boom, J.H., Rich, A., (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3611-3615.
16. Tinoco Jr., I., Davis, R.C., Jaskunas, J.R., (1968) in *Molecular Association in Biology*, Pullman, B., ed. Academic Press, New-York, p. 77.
17. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.I., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., III, Slocumbe, P.M., Smith, M., (1977) *Nature* **265**, 687-695.
18. Ehrlich, M., Marin, M., Larousse, A., Gabarro-Arpa, J., Schmitt, B., Reiss, C., (1983) *FEBS Symposium on DNA, Liblice, Folia Biologica* **30**, 105-118.
19. Bensimhon, M., Gabarro-Arpa, J., Ehrlich, R., Reiss, C., (1983) *Nuc. Acids Res.* **11**, 4521-4540.