Methods 47 (2009) 159–167

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Thermogenomics: Thermodynamic-based approaches to genomic analyses of DNA structure

P. Shing Ho *

Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, 316 MRB Building, Colorado State University, Fort Collins, CO 80523-1870, USA

ARTICLE INFO

Article history:
Accepted 12 September 2008
Available online 10 October 2008

Keywords:
Z-DNA
Nuclear factor I
Transcription regulation

ABSTRACT

The postgenomic era is all about learning about function by comparing genomic sequences within and between organisms. This review describes an approach that applies detailed thermodynamic information, as opposed to sequence motif searches, to analyze genomes (thermogenomics) for the occurrence of sequences with the potential to form left-handed Z-DNA and those that bind the eukaryotic nuclear factor I (NFI) transcriptional regulators. Such thermogenomic strategies allow us to address the questions of whether Z-DNA forming sequences can potentially function in regulating transcription of eukaryotic genes and how such function may emerge relative to other GC-rich elements, such as NFI recognition sites, to become a transcriptional coactivator.

1. DNA structures and their functions

1.1. Summary of DNA structures

It is now over 55 years since James Watson and Francis Crick [1] proposed a right-handed double-helical structure of DNA (the B-DNA structure) as a means to replicate and express the genetic information in a cell. Since that landmark paper, DNA has proven to be a highly polymorphic biomolecule, with a number of alternative structures identified, many of which are associated with specific cellular functions. For example, the four-stranded G-quartets are found at telomer ends of chromosomes and are associated with cellular senescence [2,3] and cancer progression [4,5]. However, the assumption is that the original B-form of DNA is the dominant conformation in the cell; indeed, for most biologists, all that is required to understand the biological function of DNA is the generic antiparallel Watson–Crick base paired double-helix.

Still, the plethora of non-B-DNA conformations has become exceedingly large—indeed, the historical approach of naming new DNA structures using letters of the alphabet has run into the problem that there are few letters from A to Z [6] that have not already been assigned, leaving very few letters available to name new conformations (Fig. 1). Table 1 provides a summary of some structurally interesting conformations, but is not meant to be exhaustive as there have been a number of reviews on all the various DNA structure [7,8]. Nearly all of these conformations have been characterized by X-ray diffraction studies on fibers and/or crystals, or by NMR; thus, there is plenty of detailed structural information, often times to atomic resolution, on these various DNA conformations. There is, however, much less known about how or if any of these structures contribute to genetic functions in the cell.

A very interesting case in point is Z-DNA, a rigid, left-handed variant of the standard right-handed double-helix [9]. Although it was the very first DNA structure to be crystallographically determined at atomic resolution (the atomic structure of B-DNA followed a couple of years later [10]), its potential biological function has been debated for the thirty years since its discovery [11–14]. The initial strategies to establish a biological function for Z-DNA focused primarily on biochemical and biophysical studies, including mapping its occurrence in functional chromosomes with Z-DNA specific antibodies and chemical reagents, along with attempts to isolate and identify functional proteins with specificity for the left-handed double-helix. It had been the missteps and inability during these early concerted efforts to identify "Z-proteins" that made the structure a near pariah in biology [14]. In recent years, a number of proteins have in fact been found to recognize the structure of Z-DNA in a sequence-independent manner, including proteins associated with RNA editing [15], gene transactivation [16], etc. However, it is also known that Z-DNA does not bind to proteins that either require the base-sequence information in the grooves of the right-handed helices for recognition (including, for example, transcription factors), or require the helical flexibility of B-type DNAs, such as the histone proteins that package DNA into nucleosome structures [17,18]. This latter property, in fact, has been proposed as providing a means for Z-DNA to serve as a transcriptional coactivator [19]. This review will focus on the location of Z-DNA in genomes as an example of how genomic

Abbreviations: ZDR, potential Z-DNA regions; NFI, nuclear factor I; CSF, colony stimulating factor.

* Fax: +1 970 491 0494.
E-mail address: Shing.Ho@colostate.edu.

1046-2023/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2008.09.007

and phylogenomic analyses can help to elucidate the functional role of a structure as well as its evolutionary emergence relative to other functional elements in the cell.

In 1985, before genomics and bioinformatics were catch phrases, a few postdoctoral associates in the laboratory of Dr. Alexander Rich decided that we could take a computational approach to addressing the question of whether the left-handed structure had any biologically interesting function. To do this, we developed a computer program (called ZHUNT) to study the occurrence of sequences with high propensities to form Z-DNA in genomes [20]. The idea was that if the structure has a function, then sequences with the propensity to adopt this left-handed form would localize, accumulate, or be suppressed in genomic domains with known functions. If, for example, Z-DNA were to be involved in transcriptional regulation, it should be found near markers that control gene expression, including the transcription start-site (TSS) [21], or be coupled with binding sites for transcription factors. Such a bioinformatics approach should be general and not depend on finding any specific protein that recognizes the structure. This review will discuss *in silico*, thermodynamics based approaches to identifying and locating non-B-DNA structures in genomes.

There are now many widely available tools to search for simple sequence motifs [22] and, since all DNA conformations have sequence preferences for their formation, one might expect that searching for unusual conformations would be relatively straightforward since most non-B-DNA type structures are typically characterized by simple repeating motifs. Inverted and mirror repeats, for example, are motifs that are characteristic of cruciforms and H-DNA and, therefore, algorithms that search for such sequence patterns can be used to identify the genomic occurrences of and potential functions for such structures [23–25]. Using, again, Z-DNA as the example, our experience is that simple sequence rules do not accurately identify the location or, more importantly, the probability or propensity of a sequence to adopt the left-handed conformation. For example, if we were to search for one turn of the two sequences that are known to form Z-DNA (CGCGCGCGCGCG, and CACACACACACA), statistically, we would expect to find just 6 CpG type and 60 of the CpA type sequences in 1 billion base pairs (for the human genome, which is 41.5% GC in content). We know, however, that such sequences are much more prevalent than predicted. In addition, many eukaryotic promoters are not strictly simple repeats of CpG or CpA/TpG dinucleotides (for example, promoter sequences of the rat prolactin gene [26]) and, therefore, such sequences would not be identified in a search for repeating elements, nor would we have a measure of their propensities to adopt the left-handed form relative to these standard repeating elements. Thus, this review will focus primarily on approaches that apply thermodynamic rules to search for DNA structures, particularly

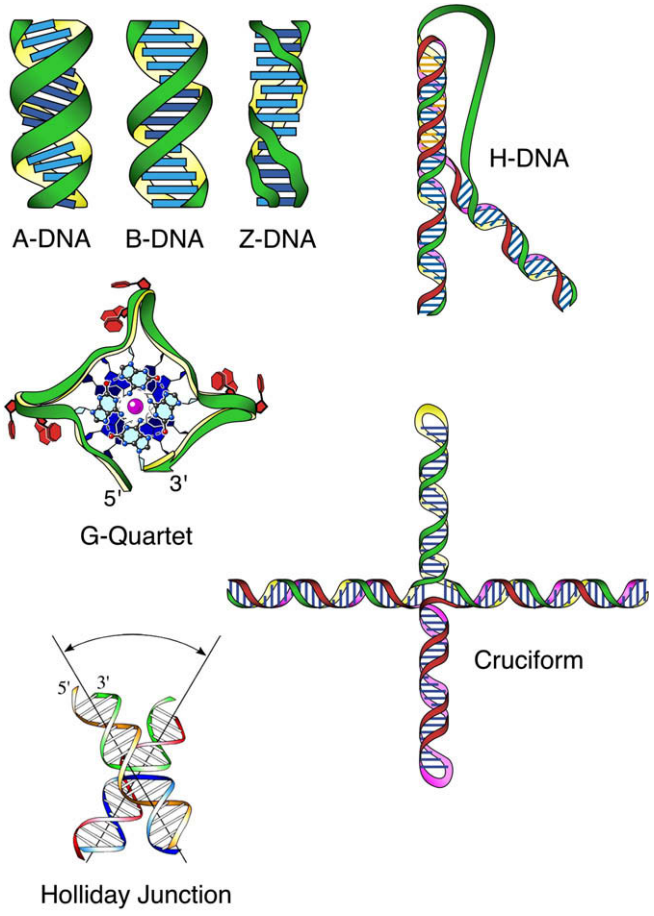

Fig. 1. Structures of various DNA conformations. Shown are the double-stranded conformations of A-, B-, and Z-DNA (ribbons trace the phosphodeoxyribose backbone, light bars represent the base pair faces in the major groove, dark bars for the minor groove), the triple-stranded H-DNA form, the G-quartet structure of the human telomer sequence [66], the extruded cruciform structure, and the Holliday junction [61] at the core of cruciforms and other recombination intermediates.

Table 1
Description of several DNA forms and their sequence dependences

DNA form	Description/potential function	Sequence dependence
A-DNA	Right-handed duplex with Watson–Crick base pairs, 11 bp/turn, 2.6 Å helical rise Function: Implicated in RNA polymerase recognition [56]	Nonalternating GC-rich; GGN, NGG and CC(C/G)
B-DNA	Right-handed duplex with Watson–Crick base pairs, 10–10.5 bp/turn, 3.4 Å rise Function: Canonical structure of DNA	All sequences
Cruciforms	Extruded DNA duplexes with 2 B-DNA stem-loops connected by a four-way junction (see Holliday junction)	AT-rich inverted repeats
"Extended"—DNA	Extended right-handed A-like duplex with Watson–Crick base pairs; 12.2 bp/turn, 3.6–3.7 Å rise Function: Intermediate in B- to A-DNA pathway; spontaneous deamination of mC in transition mutation to T [49]	Cytosine methylation in GC-rich sequences
G-Quartet	Four-stranded structure with G·G·G·G Hoogsteen type base pairs Function: Telomeric ends [57–59]	Stings of consecutive G's
H-DNA	Three-stranded helix with Watson–Crick and Hoogsteen type base triplets Function: Mutagenesis in mammalian cells [60]	Mirror repeats
Holliday Junctions	Four-stranded structure with B-type arms Function: Recombination and recombination dependent processes (DNA repair, etc.) [61]	GC-rich inverted repeats: NYC (N = A > G > C; Y = C > T)
i-Motif	Four-stranded structure with intercalated C·C+ base pairs Function: Telomer ends of human chromosomes [62–64]	Strings of C's
Z-DNA	Left-handed duplex with Watson–Crick base pairs, −12.0 bp/turn, 3.7 Å rise Function: RNA editing; transcription regulation; DNA deletion [12,65]	Alternating YR dinucleotides (Y = C > T; R = G > A)

Listed are a series of DNA structures, brief description of their structural features and potential functions, and their sequence dependency for formation.