

IEEE 7th BIBE Keynote: Promoter studies in the human genome: one perspective on an unfinished story

Mary Qu Yang

National Human Genome Research Institute,
National Institutes Health (NIH),
And Oak Ridge, DOE
yangma@mail.nih.gov

Laura L. Elnitski

Head of Genomic Functional Analysis Section,
National Human Genome Research Institute,
National Institutes Health (NIH),
Rockville MD, 20852 USA
elnitski@mail.nih.gov

The last 50 years have marked a rapid expansion in our knowledge of the genetic material known as DNA. In 1953 Jim Watson and Francis Crick published a series of papers proposing the structure and function of DNA. X-ray crystallography results from Rosalyn Franklin and Maurice Wilkins confirmed the conclusion of a double-stranded, anti-parallel helix with a sugar and phosphate backbone enclosing a stack of nucleotide bases [1-4]. In those days the discovery of the double helix structure united the interests of crystallographers, biochemists and phage geneticists, and the discipline known as ‘molecular biology’ was born [5].

The 50th anniversary of the discovery of the structure of DNA was celebrated around the world, reflecting the enormity of this scientific achievement (<http://www.nature.com/news/specials/dna50/index.html>). Yet keeping our eyes on the future, we eagerly advanced into new realms of investigation. Knowing the entire sequence of the human genome (sans a few challenging regions), the community began to speculate about integrating large-scale genomic studies in order to construct a comprehensive biological story. The ingenuity of this idea was later confirmed through the successful outcome of the ENCODE Consortium [6], sparking a new era in the study of the functional elements in genomic DNA. The endeavor integrated diverse types of biological data corresponding to functional elements in 1% of the human genome and united the interests of biologists, computer scientists, statisticians and engineers. Henceforth the discipline of ‘computational biology’ grew into a field of its own—formally defined as the hypothesis-driven investigation of a specific biological problem using computers and carried out with experimental or simulated data, with the

primary goal of discovery and advancement of biological knowledge.

One category of genomic regulatory features generating intense interest by its compelling biological relevance and unsolved questions is the promoter—the regulatory region that controls expression of a gene through a process known as transcription initiation. The earliest descriptions of functional promoter elements centered on the importance of a TATA-motif to recruit the essential RNA polymerase II molecule to the transcription start site (TSS). We now understand that the TATA-centric view of promoters represents only a minor proportion of promoters in eukaryotic (non-bacterial) cells [7]. Over the years, diverse combinations of regulatory elements have been added to the list of features capable of activating transcription in the absence of TATA-motifs.

The discovery of alternative functional elements in promoter sequences has expanded the scope of gene regulation to include how these elements could be differentially utilized. Subsets of mammalian promoters began to emerge containing initiator elements, CpG islands, downstream promoter elements and TFIIB recognition elements. Insight into the selective activation of promoters showed that TATA motifs were more likely to regulate tissue-specific expression [8] whereas CpG islands were associated with housekeeping genes having ubiquitous expression patterns [9]. Nevertheless novel functional elements in promoter sequences continue to elude detection by standard approaches.

As larger datasets emerge, the technique of subclassifying promoter sequences based on their characteristic features enables further investigation, such as the precise positioning of elements relative to the TSSs. This approach provides evidence that elusive

promoter features continue to go undetected, because after such classification, promoters containing no characterized features are found. When clustering promoter sequences by the homogeneity of their content or their function, we begin to see the emergence of biological regulatory themes. For instance the stress-response of TATA-box promoters in yeast is one well-known example of a coordinated functional response by a class of promoters [10].

As the field of gene regulation shifts from studies of individual genes to the characterization of promoter features on a species-wide scale, handling enormous sets of data will become customary. Increasingly, high-throughput experimental studies are providing a wealth of information that is useful for deducing biologically relevant themes. Assays such as ChIP-chip¹ are powerful investigative tools for revealing the presence of a protein bound to DNA. The cost and labor involved with such studies are large; however the significance of these experimental results far exceeds any other method for obtaining binding information at this scale. For example, ChIP-chip data revealed the binding of RNA polymerase II at the collection of active promoters in the cell, providing a snapshot of the inner workings of the cell [11].

Data derived from ChIP-chip assays can reveal the occupancy of promoters on a genome-wide scale, yet are collected for only one protein per experiment. In contrast, predictive approaches to finding binding sites have the ability to reveal the landscape of regulatory sites represented by numerous discrete motifs throughout the genome. However, pattern mapping and pattern discovery techniques have yet to reach their full potential as the prevailing confirmatory approach, due to limited knowledge of which patterns represent a good binding site, how accessible the DNA might be at any true site, and the number of transcription factors that have not been identified.

Another aspect of promoter research, the *de novo* identification of precise boundaries of the functional region also requires increased sophistication. By conventional means, datasets representing promoter sequences have been assembled from the sequences of DNA that surround the transcription start sites of genes. The amount of sequence used in such analyses is selected in an arbitrarily inclusive manner, for instance,

¹ ChIP-chip refers to the stages of the technique. For instance, the DNA residing in a compressed structure (Chromatin) can be extracted from the cell as fragmented sequences (~500bp) interacting with the protein of interest, through an immunoprecipitation step (IP). The final stage, in which the recovered DNA fragments are hybridized to a series of microarray chips, correlates the signal from a binding event to the positions of protein binding across the genome.

by taking roughly 550 bp upstream and 50 bp downstream of the transcription start site [12]. More recently, sets of transcripts with precise initiation sites have been produced and mapped onto their positions in the genome. This experimental technique, known as cap-trapping or CAGE [13], precisely defines TSSs by capturing all transcripts at their first nucleotide (recognized by its methylated *cap*). This cap is “worn” at the beginning of the transcript, which corresponds to the “head” or beginning of the gene. Data generated by cap-trapping assays promise to significantly advance our knowledge of the transcriptome in any given cell type, refine our knowledge of the start sites of genes, and, by inference, pave the way for promoter analyses utilizing the sequences immediately upstream and downstream of the TSSs.

In addition to better annotations of promoter locations in the cell, precisely mapped TSSs enable focused studies of the range of functional nucleotides—upstream and downstream, involved in initiation of transcription. Each and every promoter sequence associated with the ~25,000 protein coding genes in the human genome (including alternative promoters), might vary in their length of functional sequence.

Moving beyond the identification of the static location of promoters and transcription factor binding sites, knowledge of their dynamic use is at the forefront of the field of gene regulation. Recent reports of the epigenetic landscape of the DNA sequence promise a fruitful, yet challenging area of research [14]. Epigenetic modifications do not change the DNA code itself, but rather, influence the availability of the sequence to transcription factors. Epigenetic modifications can make a gene accessible and thus increase the level of its product, or make it inaccessible—effectively switching it off. In promoters, distinct acetylation and methylation modifications on histone proteins mark those promoters that are actively transcribed or silent, respectively. Epigenetic alterations can be found using the same ChIP-chip technology used to determine transcription factor binding. Although the language of the epigenetic code is largely unknown, it is recognized as a powerful mechanism in gene regulation.

The amount of research effort focused on gene regulation over the past several decades has provoked comments that there are no novel insights remaining in promoter research. Nevertheless, unique mechanistic details are emerging and are expected to continue into the foreseeable future. As paradigms of gene regulation converge into centralized themes, we can return to the accomplishments of Watson and Crick with the implicit understanding that the structure of DNA provided an

important pedestal upon which the embodiment of the genome sequence was built. Adorning this structure will be the crowning achievement of understanding how the component parts of the genomic sequence work together. Until then, our knowledge of the genome remains an unfinished story.

Acknowledgments

L.E. is supported by the Intramural Research Program of the National Human Genome Research Institute, US National Institutes of Health. Special thanks to Dr. Edwin Jacox for helpful feedback during the preparation of this manuscript.

References

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171: 737-738.
2. Wilkins MH, Stokes AR, Wilson HR (1953) Molecular structure of deoxypentose nucleic acids. *Nature* 171: 738-740.
3. Franklin RE, Gosling RG (1953) Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature* 172: 156-157.
4. Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. *Nature* 421: 400-401; discussion 396.
5. Strasser BJ (2003) Who cares about the double helix? *Nature* 422: 803-804.
6. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816.
7. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res* 11: 677-684.
8. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, *et al.* (2006) The complexity of the mammalian transcriptome. *J Physiol* 575: 321-332.
9. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196: 261-282.
10. Zanton SJ, Pugh BF (2004) Changes in genomewide occupancy of core transcriptional regulators during heat stress. *Proc Natl Acad Sci U S A* 101: 16843-16848.
11. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature* 436: 876-880.
12. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res* 13: 308-312.
13. Shimokawa K, Okamura-Oho Y, Kurita T, Frith MC, Kawai J, *et al.* (2007) Large-scale clustering of CAGE tag expression data.
14. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. *Cell* 128: 669-681.



Dr. Laura L. Elnitski received her Ph.D. from Pennsylvania State University and was a recipient of NIH Post Doctoral Fellowship. Dr. Elnitski is a molecular and computational biologist who uses experimental and bioinformatic methods to discover noncoding functional elements in the human genome and has made seminal contributions to the field. She is on the editorial board of *Genomic Research*. She is currently Head of NIH/NHGRI Genomic Functional Analysis Section