MATHEMATICAL MODELING AND THE HUMAN GENOME

Hilary S. Booth
Australian National University, Australia

Keywords: Human genome, DNA, bioinformatics, sequence analysis, evolution.

Contents

Summary

Mathematical techniques currently used to model the human genome are discussed. Human DNA encodes all of the functional elements necessary for the development and maintenance of a human being within an external environment. As such, mathematical models of the human genome attempt to not only describe the evolution of DNA but also the structure and function of the biological elements encoded by the genome.

Background into the sequencing of the human genome and the "central dogma" of molecular biology is given in the first section. Models of the evolution of DNA are then discussed. The third section on modeling genes deals with gene-finding using hidden Markov models and the measurement of similarity between genes.

1. Introduction: The Human Genome

Within the nucleus of every cell of every human, long coils of DNA (deoxyribonucleic acid) form the chromosomes that contain encoded information necessary for the human being to develop, within a changing external environment, from a fetus to a child and as an adult, to reproduce and eventually to die.

The genetic information in DNA is encoded in a sequence of the four chemical bases or nucleotides (adenine, thymine, guanine, cytosine) abbreviated as A, T, G and C. The coding DNA is accompanied by a second complementary strand that is fully determined by the nucleotides in the coding strand (A $\Leftrightarrow$ T; G $\Leftrightarrow$ C). The two strands join together to form the double helix whose structure was discovered in the 1950's, for which

Watson and Crick were awarded the Nobel Prize in 1962. DNA is also present in the *mitochondria* (organelles within the cell) and this information is passed exclusively down the maternal line.

The human genome is the sum of all DNA (chromosomal and mitochondrial) in the cells of a human. The genome includes genes i.e. those sections of DNA that contribute to a function, which in turn determine physical appearance, certain behavioral characteristics, how well the organism combats specific diseases and other characteristics. The genome also includes mysterious regions of unknown function, often referred to as "junk DNA", since it is widely agreed that much of the information contained in the human genome is superfluous.

The entire human genome consists of approximately 3,000,000,000 pairs of bases, and is thought to contain around 30,000 genes. The Human Genome Project was a global, long-term research effort (officially begun in 1989) to determine the sequence of the chemical bases that make up human DNA, to identify the genes and make their sequences available for further biological study. The finished sequence announced in April 2003 covers about 99 percent of the human genome's gene-containing regions, which have been sequenced to an accuracy of 99.99 percent. The DNA in the human genome contains the information necessary to construct gene products such as proteins, which are made on the basis of mRNA.

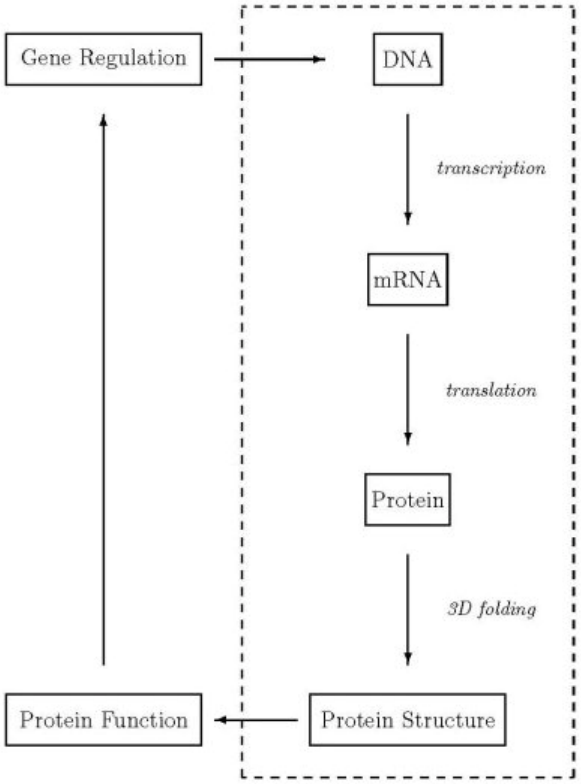

Figure 1: The central dogma of genetics

The way in which these products are constructed depends upon the information encoded in the DNA, and minor differences sometimes affect biological attributes and functions

resulting in differences between individuals. Mathematical models of the human genome attempt to describe this process at various levels. Ultimately, the aim of genetics is to shed light upon the relationship between genotype (the DNA) and phenotype (the physical characteristics) of an individual.

In Figure 1 we show what is known as the "central dogma" of molecular biology. The central dogma is a simplification of the complex biological system of interactions that eventually results in protein production.

Once a gene has been activated by the gene regulatory network it is expressed in the cell i.e. the gene's DNA is transcribed into mRNA which is in turn translated into a protein sequence. The protein folds into a 3D structure. In response to the developmental requirements of the cell and the demands of the external environment, the proteins perform functions. The differences between the ways in which proteins perform these functions will result in different phenotypes. The phenotype then is a result of both the genotype and the environment.

2. Modeling DNA

All models of the human genome attempt to describe aspects of the molecular biology encoded in the DNA. The structure summarized (in simplified form) in the central dogma can be modeled mathematically at any level of detail. Models that relate the different levels are often required, and some levels will be omitted for simplicity. The mathematical model that is used will depend upon the biological questions the model is to address and the available data.

Models of evolution for example approximate the gradual change in DNA due to mutation, from generation to generation. Different evolutionary models incorporate various simplifying assumptions. While describing the evolution of DNA at the molecular level, the different models also incorporate hypotheses regarding the evolution of the phenotype.

Hidden Markov models (HMMs) proved to be one of the most successful methods of locating genes and other important features of the human genome. Based upon known features, the HMM is "trained" to recognize regions of DNA sequence that are likely to contain similar features. HMMs are able to predict gene structure more accurately than other models due to their ability to incorporate background genomic sequence composition as "hidden states" of the Markov model.

One of the most useful advances in bioinformatics has been the development of methods of measuring the similarity between DNA sequences leading to the ability to find homologous (related) genes in the public databases. The most successful method of measuring similarity is based upon a (null) random walk model. The similarity of two sequences is assessed against the null model and it is determined whether or not two randomly generated sequences of the same length are likely to have been as similar. (See Section 3.)

2.1 Evolutionary Models

Ideally, a useful mathematical model of the human genome should be able to model its evolution throughout time. For computational purposes, a DNA sequence is considered to be a string of characters taken from the four-letter alphabet A,T,G,C. The process of evolution essentially consists of changes (mutations), insertions or deletions occurring in this string from generation to generation. These changes are usually assumed to occur at random. If a mutation results in a genetic variation that results in the early death of the organism, then that mutation will not be carried on to the next generation. In this way, natural selection determines those mutations that do not survive the evolutionary process.

The Neutral Theory of Molecular Evolution, a slightly modified version of the Darwinian theory of evolution by natural selection, is now widely accepted as the best model of the evolutionary process at the molecular level. The neutral theory asserts that the great majority of mutations that survive throughout evolution are selectively neutral. Most of the changes in DNA throughout the evolutionary process are caused by the random fixation of selectively neutral or nearly neutral mutants in the species, rather than by positive Darwinian selection. This implies that there are no important survival advantages or disadvantages associated with particular alleles and that genetic drift, rather than natural selection, dominates the evolutionary process. This does not mean that mutations, when they occur, are all neutral, or that the genes themselves are unimportant. On the contrary, it is thought that many mutations are deleterious to the organism, and thus are unlikely to remain in the population for long. Only those mutations that do not have a harmful effect will survive long enough to exist in a significant proportion of the population. The alternative (neo-Darwinian) viewpoint is that advantageous mutations, while perhaps exceedingly rare, do play a major role in evolution, and that most of the polymorphism at the molecular level can best be explained by natural selection.

The neutral theory is supported by DNA evidence that shows that molecules or parts of one molecule that are subjected to weaker functional constraints evolve faster (more nucleotide substitutions per generation) than those that are subjected to stronger constraints. It has been demonstrated for example that in coding regions, the third codon (which is often interchangeable with another nucleotide) evolves more rapidly than the first or second codon (at which position the change of a nucleotide would result in a different amino acid in the protein). In other words, nucleotide substitutions that are selectively neutral occur much more often than substitutions that cause a change in the molecule encoded by the DNA.

Note that the environment plays a role here too. Changes in the environment may affect the survival of different phenotypes, often in unpredictable ways. For example, the sickle-cell anemia mutation in human hemoglobin is disadvantageous in general but advantageous in malaria-infected populations, and the corresponding haplotype occurs more frequently under these conditions. Even with the environmental effect taken into account, it is still generally true that fewer mutations per generation are expected in functionally important regions of the genome, and more mutations in regions less crucial to the survival of the organism. Highly conserved gene sequences tend to be functionally important and related to each other and this observation is the basis of most phylogeny (tree-building) and sequence comparison techniques.

An accurate evolutionary model would ideally incorporate regions of highly conserved DNA (corresponding to the "essential" regions) interspersed with regions that evolve more freely (in regions that are either non-coding or less important for the structure and function of the gene product). However there are difficulties associated with modeling this situation accurately. Firstly, we do not fully understand the ways in which DNA, RNA and proteins interact in order to function. Furthermore, the more features we add to our model, the more difficult we make our computational problem, and given the length of the human genome this latter point is extremely important. For these two reasons, it is sometimes necessary to use simplified models of nucleotide substitution when modeling the evolution of the human genome. A more detailed description using hidden Markov models incorporates the different features of the coding regions and was developed to model genes more accurately. (See Section 3).

-
-
-

Bibliography

Altschul, S.F., Gish W., Miller W., Myers E. W. & Lipman D. J. (1990) Basic local alignment search tool. *J.Mol.Biol.* 215, 403-410.

C. Burge and S.Karlin (1997) Prediction of Complete Gene Structures in Human Genomic DNA *J. Mol. Biol.* 268, 78-94.

Ewens, W. J. & Grant, G. R. (2000). *Statistical Methods in Bioinformatics: An Introduction*, Springer, New York, USA.

Karlin, S. and S.F. Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. USA*, 87, 2264-2268, 1990.
Lander E.S. et al (2001) Initial Sequencing and Analysis of the Human Genome. *Nature* 409:860-921.

M. Kimura (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge

Needleman S. B. & Wunsch C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins *J. Mol. Biol.* 48, 443-453.

R. M. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis.*Cambridge University Press, Cambridge, UK, 1998.

Smith T.F. and Waterman M.S. (1981). Identification of common molecular subsequences *J. Mol.Biol.*147:195-197.

Venter J.C. et al (2001) The Sequence of the Human Genome. *Science* 291: 1304-1351.