Genefinding

By:
Andrew Hughes

Genefinding

By:
Andrew Hughes

Online:
< http://cnx.org/content/col10205/1.1/ >

CONNEXIONS

Rice University, Houston, Texas

Collection structure revised: June 17, 2003
PDF generated: February 4, 2011
For copyright and attribution information for the modules contained in this collection, see p. 25.

Table of Contents

Chapter 1

Introduction to Genefinding[1]

Just under fifty years ago two discoveries were made that have changed, or will soon change, the way the human race understands and experiences life. In 1947 Bardeen, Brittan, and Shockley discovered the transistor effect, an achievement for which, in 1956, they received the Nobel prize. A few short years later, Jack Kilby at Texas Instruments (1958) and Robert Noyce at Fairchild Camera (1959) made the breakthroughs from which the integrated circuit (multiple transistors on one substrate) was later developed.

[1] This content is available online at <http://cnx.org/content/m11315/1.8/>.

The first transistor

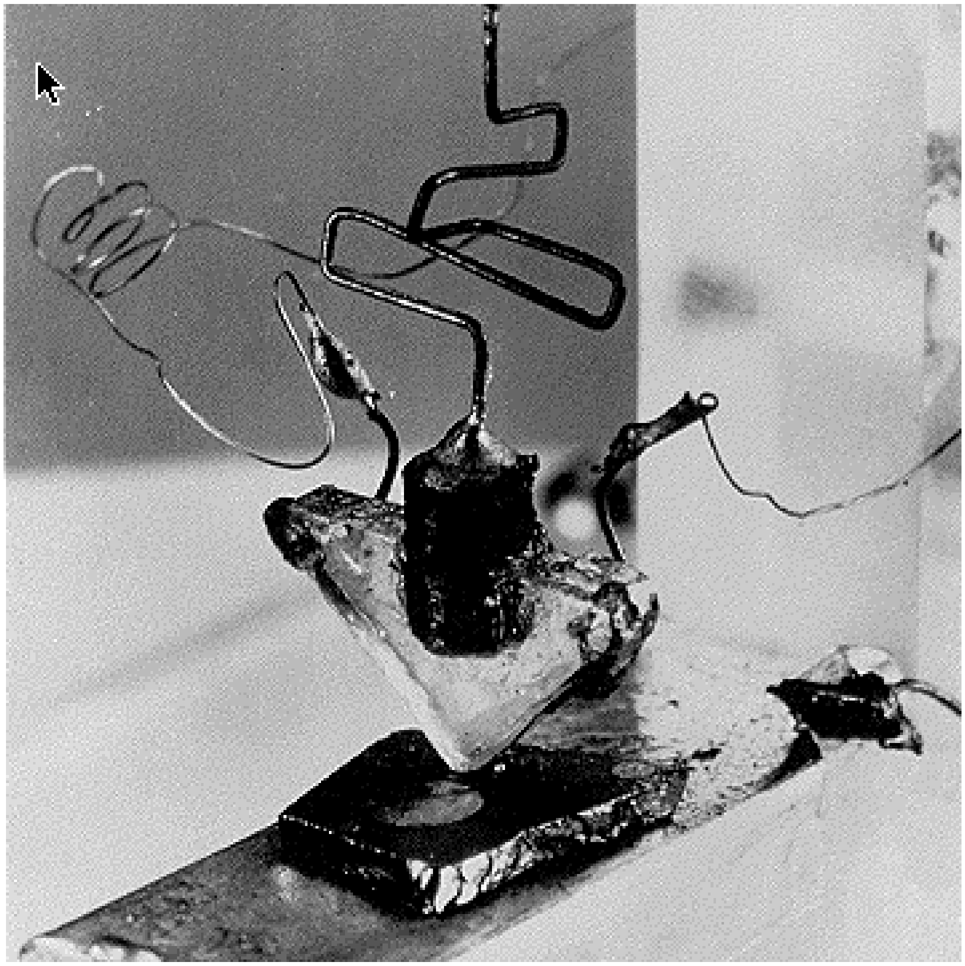
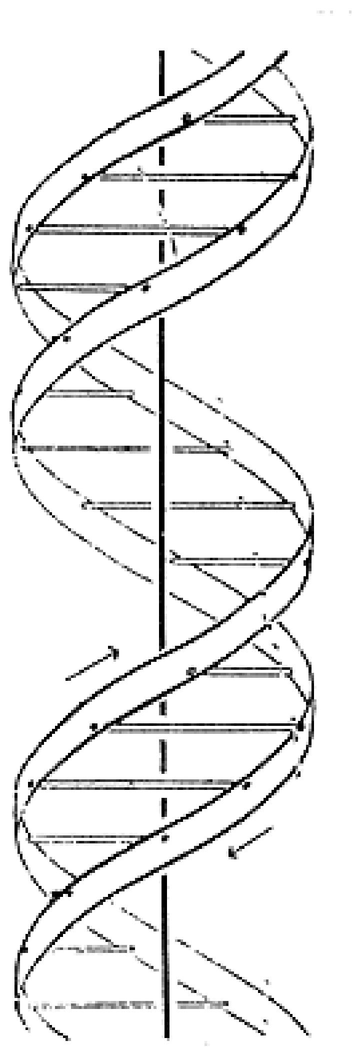


Figure 1.1: This is a picture of the first transistor invented at Bell Labs by Bardeen, Brittan, and Shockley in 1947.

In 1953 Watson and Crick unlocked the structure of the DNA molecule and set into motion the modern study of genetics. This advance allowed our study of life to transcend the wet and dirty realm of proteins, cells, organelles, ions, and lipids, and move up into more abstract methods of analysis. By discovering the basic structure of DNA we had received our first glance into the information-based realm locked inside the genetic code.

The structure of DNA



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Figure 1.2: This is the original figure depicting the structure of DNA that was published in Nature in 1953.

For almost fifty years genetics and computer science coexisted side-by-side without much interaction. Computer science is a discipline of abstract data, concerned with its manipulation, management and analysis. Genetics was empirical in nature and experiment driven; analyzing large amounts of data was simply not a problem geneticists had to contend with.

Historically, for geneticists, the amount of time required to generate data has been much greater than the amount of time required to analyze the resultant data. Genetics experiments that necessitate weeks, months, or even years to bring to successful fruition can generate as little as a few radiographic films in data. After a successful experiment, a triumphant geneticist would leisurely (between writing grants and teaching classes) scour the data as a hungry dog would attack a leftover steak, analyzing every available nook and cranny, leaving no promising datum untouched. In the past, generating good experimental data was analogous to searching for gold in that it was scarce and highly-appreciated if it was found.

One problem geneticists are not accustomed to contending with is an over-abundance of information. In some ways, the reluctant pace of incoming information has been a good thing for genetics; the interpretation of genetics data can be a challenging endeavor and the relatively slow tempo of its acquisition has allowed ample time for researchers to fully appreciate the implications of each small piece as it arrived.

Rosalind Franklin

Figure 1.3

Human beings and computers have divergent and complimentary abilities. Computers are intrinsically beasts of information; they deal with pure abstract data, ones and zeros. Relative to humans, computers excel at manipulating large amounts of data, performing numerous calculations quickly, and analyzing large, multi-dimensional data sets. Humans, by contrast, are physically rooted in nature and have a proclivity for higher abstract thinking, long-term planning, and assimilating noisy or incomplete information. We are flexible and adaptable where computers are efficient and rigid.

As the powers of computers has developed and matured, the manner in which we use them has correspondingly evolved. Initially computers insinuated themselves into our lives because of their ability to quickly perform large numbers of simple calculations, and because they could be used to efficiently store large amounts of information. Used as such, they were essentially glorified calculator-filing-cabinets.

Watson and Crick

Figure 1.4

Today however, we can go well beyond this simple understanding of our relationship with computers as experimental tools. This changing dynamic is especially evident in, and necessary to, the emergent field of bioinformatics where successful realization of the presented challenges requires both the computer's ability to analyze large and complex data sets; and the human ability to initially generate the data as well as to interpret the computer analysis of the data. Computers should be viewed as tools to extend our vision into the abstract realms of data analysis, and this improved sight should improve our efficiency in the laboratory.

This type of symbiosis is commonplace today. An example scenario might be as follows: a researcher isolates a novel gene of interest and has it sent off to be sequenced. When the researcher receives the sequence in the mail a few days later, the researcher then loads the sequence into the BLAST search engine looking for known homologues. If a homologue exists, either in the same species or in another, related, species, this information can be used to predict the possible functions the gene might have. Alternatively, the researcher might want to isolate where the gene resides in the genomic DNA. Before whole-genome sequences were available, this was a very laborious and difficult process involving time intensive restriction mapping techniques. Today, the process has been greatly simplified. To find the gene's location in the genomic DNA the researcher would almost certainly begin with a BLAST search of the organism's genome (if available, or a closely related organism if not). The search would return a list of candidate sequences, and their locations in the genome, that could then be checked experimentally for identity with the gene of interest. Furthermore, a successful BLAST search might not only reveals the exact location of the gene of interest, but also any closely related genes as well (the latter being a great advantage of genomic searching versus earlier experimental gene isolation techniques).

When compared to prior techniques, a successful BLAST search is highly efficient and also returns a much greater wealth of data. Unfortunately, the BLAST search is not the end of the process. The results of the search should be viewed as candidates that must be experimentally verified in the lab before any final conclusions can be drawn about their true nature.

Maurice Wilkins

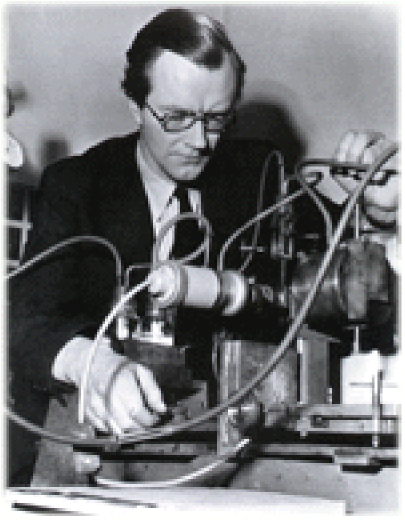


Figure 1.5

Another specific example of this type of human/computer interface can be found in the analysis of the experimental finding that 3.3% of the human genome aligns to multiple regions of the mouse genome in whole-genome BLASTZ alignments (Birney et. al. 2003). The implication of this is that outside, higher-order, human knowledge must be brought to bear on the problem of identifying the most significant alignments when multiple alignments are found. Another example that demonstrates the necessity of meaningful interaction between computer analysis and human understanding is the observation that only one third of the genome under purifying selection actually codes for protein expression (Flicek et. al. 2003). This result comes from a comparative alignment of human-mouse complete genomic sequences. The most basic implication of this is that any attempt at gene-prediction via whole-genome alignment is going to generate large numbers of false-positives because of conserved non-coding and non-regulatory regions.

In these examples we can see how experimental evidence leads to computer analysis which is then used to direct subsequent experiments. The cyclical nature of our interaction with the two search spaces, the physical and the informational, is becoming increasingly apparent as the two disciplines mature. Human exploration of the wet and chaotic physical world should direct and be directed by the computer-facilitated human exploration of the ethereal information space, which was itself generated by prior experimental insight and abstract thought. In reality, both investigative systems are indirect means of increasing our understanding of the same physical phenomena as validated by the reproducible utility of gained information when applied to either / both systems.

The inventors of the transistor

Figure 1.6: Dr. John Bardeen, Dr. Walter Brattain, and Dr. William Shockley discovered the transistor effect and developed the first device in December, 1947, while the three were members of the technical staff at Bell Laboratories in Murray Hill, NJ. They were awarded the Nobel Prize in physics in 1956.

So what, then, are the goals of genefinding as a subset of bioinformatics? Simply put, the goal of genefinding is to locate protein coding regions in unprocessed genomic DNA sequence data. In reality however, pinpointing the mere location of a gene is part of a much larger challenge. The eukaryotic gene is a complicated and highly studied beast composed of a variable multitude of small coding regions and regulatory elements hidden amidst tens of thousands of base pairs of intronic and non-signal DNA. In order to accurately predict gene locations we must first understand how the different functional components interact to create the dynamic and complex phenomena we have come to understand as 'a gene'.

Thus genefinding is something of a misleading misnomer: in order to find genes we must first understand the content and structure of the signal the genes present to the cell's genetic machinery, and in doing this we must answer much broader questions than the seemingly facile question, "Where are the genes?" The goal of genefinding then is not simple gene prediction, but accurate modeling of the signal genes present to the cell. Furthermore, because such information does not exist in a vacuum separate from it's interpretation, implicit in the assumption of the ability to model the genetic signal is a furthering of our capacity to understand the deciphering of the genetic signal and our understanding of the inner workings of the cell itself.

Chapter 2

Genetics Background[1]

The most naive picture of the eukaryotic genome is a long string of linear DNA balled up somewhere inside the cell. This formulation fails on several important grounds: first, although DNA is a linear molecule, it is not necessarily accessed in a linear fashion; second, DNA has a very significant secondary structure, it is not simply balled up at random; and third because DNA does not act in isolation, it is immersed in the context of the cell's nucleus where numerous proteins and epigenetic processes interact with the DNA to regulate gene expression.

Let's begin by discussing the in vivo structure of DNA in a typical eukaryotic cell. A molecule of DNA is composed of two antiparallel and complimentary strands of deoxyribonucleicacid. Antiparallel means that the two strands have opposite chemical polarity, or, stated another way, their sugar-phosphate backbones run in opposite directions. Direction in nucleic acids is specified by referring to the carbons of the ribose ring (ribose is a sugar) in the sugar-phosphate backbone of DNA. 5' specifies the the 5th carbon in the ribose ring, counting clockwise from the oxygen molecule, and 3' specifies the 3rd carbon in the ring. Direction of, and in reference to, DNA molecules is then specified relative to these carbons. For example, transcription, the act of transcribing DNA to RNA for eventual expression, always occurs in the 5' to 3' direction. Nucleic acid polymerization cannot occur in the opposite direction, 3' to 5', because of the difference in chemical properties between the 5' methyl group and the 3' ring-carbon with an attached hydroxyl group.

[1] This content is available online at <http://cnx.org/content/m11320/1.2/>.

DNA Helix

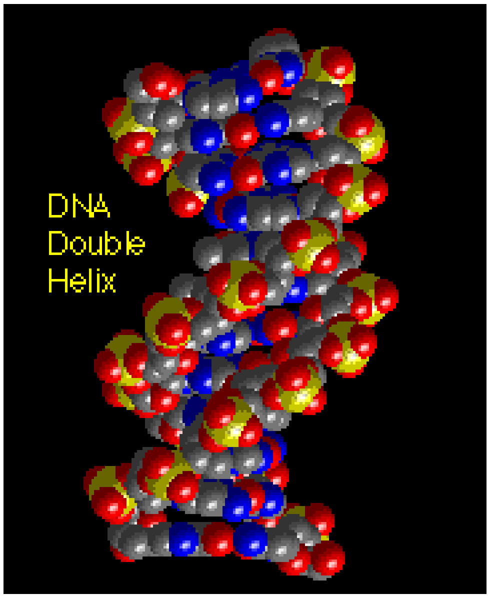


Figure 2.1

The basic structure of DNA can be divided into two portions: the external sugar-phosphate backbone, and the internal bases. The sugar-phosphate backbone, as its name implies, is the major structural component of the DNA molecule. It is the external portion of the DNA molecule because it is highly polar, and thus hydrophillic (meaning it likes to be immersed in water). Correspondingly, the interior bases of the DNA molecule are non-polar and hydrophobic. This duality has a very stabilizing effect on the overall structure of the DNA double helix: the hydrophobic core of the DNA molecule 'wants' to be hidden inside the sugar-phosphate backbone which acts to isolate it from the polar water molecules; thus there is a strong hydrophobic pressure gluing two molecules of DNA together.

There are four bases in DNA: adenine (A), guanine (G), thymine (T), and cytosine (C). In RNA uracil (U) is found in place of thymine (T). Inside a DNA molecule these bases pair up, A to T and C to G, forming hydrogen bonds that further serve to stabilize the DNA molecule. Because the interior bases pair up in this manner, we say the DNA double helix is complimentary. It is the sequence of these bases inside the DNA molecule that we refer to as the genetic code.

DNA Structure

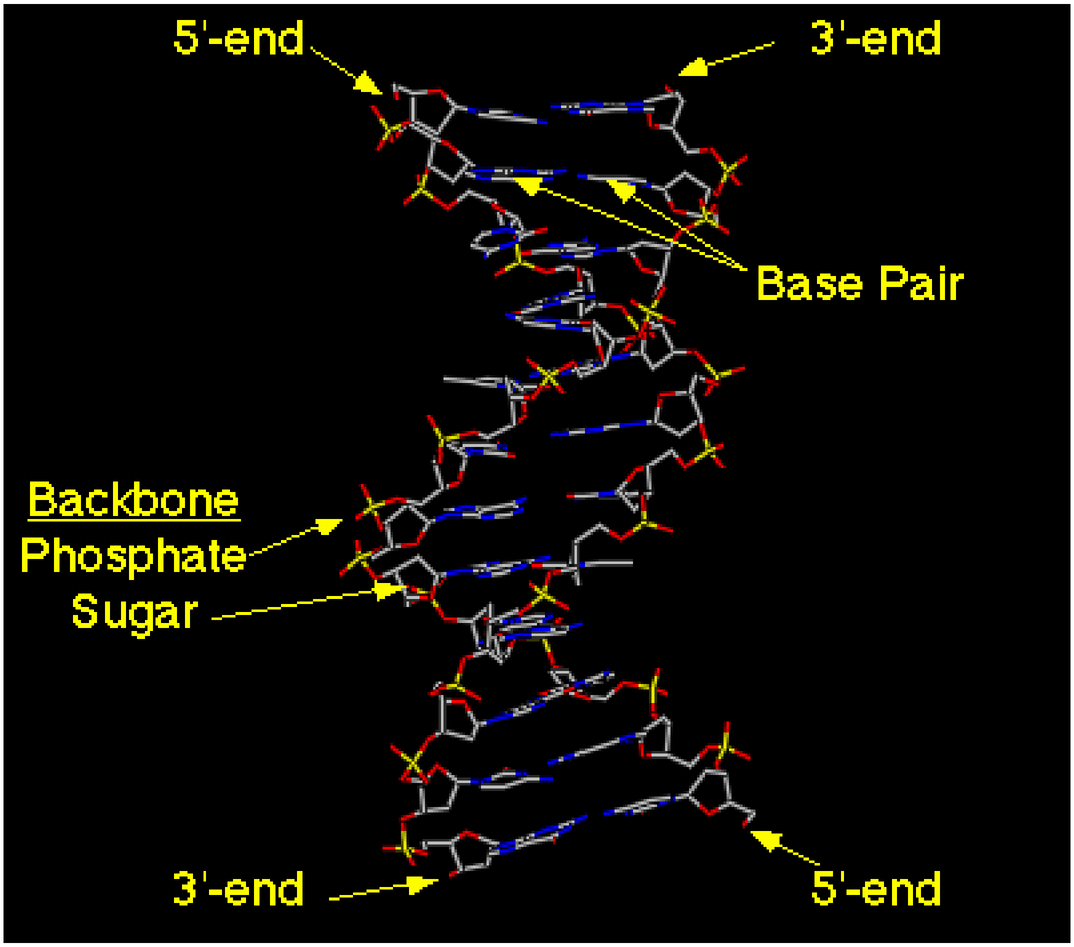


Figure 2.2

At this point we now have a good picture of the chemical structure of the DNA molecule, now we need to begin placing it in the context of the cell. A typical eukaryotic chromosome contains from 1 to 20 cm of DNA. However, during metaphase of mitosis and meiosis, this DNA is packaged in a chromosome with a length of only 1 to 10 um. How is this amazing density achieved inside the cell?

DNA in the cell exists packed into a dense and regular structure called chromatin. Chromatin is composed of DNA, proteins, and a small amount of RNA. The proteins found in chromatin largely consist of histones, a basic protein which is positively charged at neutral pH, and nonhistone chromosomal proteins which are largely acidic at neutral pH. Histones have been highly conserved in all eukaryotes. There are five major histone types, called H1, H2a, H2b, H3, and H4, and which exist in specific molar ratios within the chromatin. Histones bind together with the DNA to form the basic structural subunit of chromatic, small ellipsoidal beads called nucleosomes which are around 11nm in diameter and 6nm high. Each nucleosome contains 146 nucleotide pairs which wrap around the histon protein complex 1 and 3/4 turns. The nucleosome complexes give the DNA molecula a packaging ratio of 6.

Histones

Image not finished

Figure 2.3

Beyond the nucleosome, there are two more levels of structural packaging. The second level of packing is the coiling of the nucleosome beads into a helical structure called the 30 nm fiber that is found in both interphase chromatin and mitotic chromosomes. This structure increases the packing ratio to about 40. The final packaging occurs when the fiber is organized in loops, scaffolds and domains that give a final packing ratio of about 1000 in interphase chromosomes and about 10,000 in mitotic chromosomes.

One important note is that DNA is not always packed into the super-dense chromosome structures evident during mitotic and meiotic replication. During interphase, or the general not-currently-reproducing phase of the cell where most of a cell's work is done, the chromatin, while still highly dense, is about 1/10 as dense as during cellular replication. This is important because it is believed that the highly-dense chromatic structure of DNA sterically inhibits transcription and thus gene expression. In order for genes to be expressed the chromatin structure must be relaxed so that the transcriptional proteins can gain access to the DNA molecule.

Now that we have a good grasp on the basic structure of DNA as a molecule, as well as in vivo, lets move on to the mechanisms of gene expression. The Central Dogma of genetics is: DNA is transcribed to RNA which is translated to protein. Protein is never back-translated to RNA or DNA, and except for retroviruses, DNA is never created from RNA. Furthermore, DNA is never directly translated to protein. DNA to RNA to protein.

DNA is the long term, stable, hard-copy of the genetic material; by way of analogy it is similar to the information on a computers hard-disk drive. RNA is a temporary intermediary between the DNA and the protein making factories, the ribosomes. To further extend our computer analogy, RNA could be compared to information in a cache, in that the lifetime of RNA is much shorter than that of either DNA or the average protein, and that RNA serves to carry information from the genome, located in the nucleus of the cell, to the ribosomes, which are located outside of the nucleus either in the cytosol or on the endoplasmic reticulum (which is a large set of folded membranes proximal to the nucleus that help manufacture proteins for extra-cellular export). To complete our analogy, proteins could be viewed as the programs of the cell. They are the physical representation of the abstract information contained within the genome. However, one caveat is that RNA does have some enzymatic activity and has other functions besides ferrying messages between the DNA and the ribosomes.

Transcription is the process of creating RNA from DNA. Transcription is also the point at which most of the regulation of gene expression occurs and because of this it is a very complex process, especially with regard to its initiation. To say that DNA is transcribed to RNA is a nice (over)simplification, but we need to delve a little deeper into the details to really appreciate what is going on during transcription. A more complete view of transcription includes five steps: 1) transcription of DNA to pre-mRNA, 2) a 7-methyl guanosine cap is added to the 5' end of the transcript, 3) a poly(A) tail is added to the 3' end of the transcript, 4) the introns are spliced out of the pre-mRNA, which finally yields, 5) the mRNA transcript proper.

Because the first step, the initial transcription of DNA to pre-mRNA, is the most involved, I am going to hold off on discussing it for a moment and expand on steps 2-5 first. (2) The addition of the 5' 7-MG cap is important for two reasons: the 5' caps are recognized by protein factors that initiate translation, and it also helps protect the transcript from nucleases. Nucleases are very common in the cell and because of this unprotected RNA has a very short half-life inside the cell. Nucleases are actually so common that working with RNA in the laboratory can be quite difficult because the samples have a tendency to disintegrate into useless bits. (3) The poly(A) tails are formed in a two step process: an endonulcease cleaves around 1000-2000 non-coding bases from the 3' end of the pre-mRNA transcript and then poly(A) polymerase adds 20-200 AMP molecules to the 3' end of the transcript. The poly(A) tail is important in the cellular transport of the mRNA transcript and, like the 5' cap, also helps to stabilize the mRNA transcript.

Once the 5' cap and the poly(A) tail have been added, only one step remains for the pre-mRNA transcript to be complete and graduate to mRNA status: splicing. Eukaryotic genes contain two types of transcribed regions: introns and exons. Exons are the regions of the genome that contain actual coding information. Introns are non-coding, meaning that intronic sequences are never translated to protein, in fact they are never included in the final processed mRNA transcript. Splicing is the process of removing introns from the pre-mRNA transcript to produce an exon-only mRNA molecule, which is then shipped off for translation. Generally, eukaryotic mRNAs are considered to monogenic. However, up to one fourth of the transcripts in C. elegans have been show to be multi-genic (i.e. they contain exons from multiple genes).

A further complication of the splicing process is that mRNA can undergo alternative splicing. To illustrate this let's imagine a gene that has 3 exons and two introns. From this gene, three different final transcripts are possible. In all transcripts the two introns are going to be removed, however, the cell can combine the exons however it wants as long as the original order is maintained. This means that for this example the possible mRNA transcripts include: Exon1-Exon2, Exon1-Exon3, and Exon1-Exon2-Exon3; however, Exon3-Exon1 is not possible because the exons are out of order.

An interesting side note is that some introns are capable of self-splicing, that is they can politely remove themselves without the intervention of any proteins. This is significant mainly because it is a significant counter example to the idea that RNA is an inert transcript and action is soley the domain of proteins. RNAs should really be viewed as having both enzymatic properties and abstract information-carrying ability. Because of this many people believe that RNA was the original genetic molecule and that DNA and proteins evolved later in the game.

Alternative splicing is a very important and powerful tool. To understand the benefit alternative splicing gives the cell we need to understand something about proteins. Proteins can be understood as containing modularized functional units. These functional units can be active sites on enzymes, large structural motifs such as beta-sheets or alpha-helices, or motifs that direct the eventual destination of expressed proteins. A good example of an alternatively spliced pre-mRNA transcript is the mouse IgM immuoglobulin transcript. IgM exists in two forms: excreted and membrane bound. These two forms of the protein differ in the only in the C-terminus: the secreted protein has a secreted terminus motif while the membrane-bound protein has a C-terminal membrane anchor region. Both products come from the same pre-mRNA, but alternative splicing includes either the terminal exon that creates the excreted form of IgM or the membrane-bound form of IgM.

This is a good time to take a step back from our discussion, take a deep breath, and summarize what we have covered so far. (1) DNA exists as a double stranded helix that is both complimentary and antiparrallel. (2) DNA in vivo exists in a very compact and regular structure of nucleosomes, 30nm fibers of braided nucleosomes, and loops of fibers. (3) The central dogma of genetics: DNA is transcribed to RNA, which is then translated to proteins. (4) DNA is the stable, long-term form of genetic information. (5) RNA is (mostly) an intermediary between DNA and the protein-making-factories, ribosomes. (6) RNA transcription is not nearly as simple as the central dogma might lead you to believe. Which leads us to the point I put off earlier: how is transcription initiated in the eukaryotic genome?

Chapter 3

Genomic Data Sets[1]

3.1 Introduction to the available data sets

The amount of whole-genomic data available is mountainous and growing at wholly un-geologic rate. Currently, over 1000 whole-genome data sets are either completed or in progress (whole genomes are considered 'finished' when they contain less than one error per 10,000 base-pairs).This amazing (and daunting) source of information includes genomes from bacteria, archaea, eukaryotes, as well as viruses and organelles. In the past four years alone, entire genomic sequences from C. elegans, D. melanogaster, H. sapiens, F. rubipes, A. gambiae, M. musculus, C. briggsae, R. norvegicus, A. thaliana, and C. intestinalis have been, or are nearly, completed (Birney et. al. 2003). This data set represents a sizeable portion of the commonly studied metazoan eukaryotes. Another current example of the high-throughput power of modern sequencing facilities is the fact that within weeks of the isolation of SARS, a preliminary genomic sequence was available. As automated high-throughput genome sequencing techniques continue to progress, more and more data sets of higher and higher quality will become available. Eventually we may even progress beyond a species-based view of the genome to whole genome sequencing of individual organism. The information is quickly outstripping our ability to analyze it; we need to develop sophisticated and sensitive informational analysis tools to apply to this new wealth of information.

[1] This content is available online at <http://cnx.org/content/m11317/1.3/>.

3.2 Arabidopsis thaliana

Arabidopsis thaliana

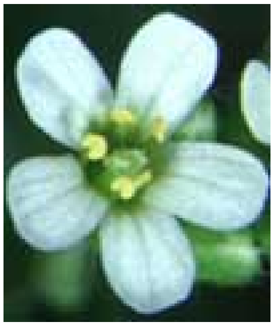


Figure 3.1

Arabidopsis is the model for plant genetics research. It is a flowering plant and a member of the mustard family; its advantages as a research model include: short generation time, small size, large number of offspring, and relatively small nuclear genome. The genome was sequenced in 2000 by The Arabidopsis Genome Initiative (Nature 14 Dec. 2000). The genome has five chromosomes and a total size of 125mb. The Arabidopsis Genome Initiative in its original analysis predicted a total of 25,498 genes; this is much larger than both C. elegans (19,000) and Drosophila (13,601) and is in the range of the estimated number of genes for H. sapiens. The average gene length is around 2000bp with the average exon being 250bp in length (∼5 per gene). The average intron is 180bp in length.

3.3 Caenorhabditis elegans

C. elegans

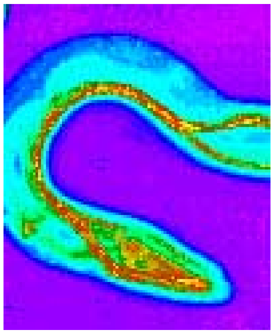


Figure 3.2

C. elegans was the first multicellular organism (it's a worm) to be completely sequenced and the second eukaryote (to yeast) to be sequenced. The genome was sequenced by The C. elegans Sequencing Consortium in 1998 (Science, 11 Dec. 1998). Before C. elegans the only other genomes to be sequenced were those of some viruses, bacteria and a yeast. The 97Mb sequence contains 19,099 predicted protein-coding genes (GENEFINDER was used to predict genes). The genome has 5 chromosomes plus an X chromosomes. Each gene has an average of 5 introns. 27% of the genome resides in predicted exons (this is much higher than human's ∼5%) and 26% of the genome resides in predicted introns. GC content in the genome is remarkably constant across all of the chromosomes (36%). Relative to higher-order metazoan eukaryotes, especially as compared to vertebrates, C. elegans presents a clean genome with a low level of repeat sequences or other low complexity regions (although they definitely do exist, ∼6%).

3.4 Drosophila melanogaster

D. melanogaster

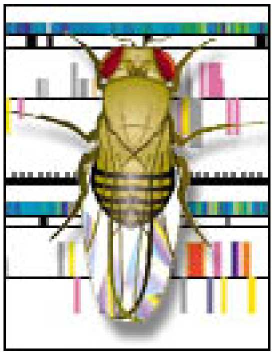


Figure 3.3

The drosophila (fruit fly) genome is 180Mb in size and contains approx. 13,600 genes (Genie and Genescan were used to predict genes). The somewhat smaller C. elegans genome actually contains more genes than the Drosophila genome, although the functional diversity between the two species appears to be very similar. The Drosophila genome was published in March of 2000 (Science, 24 March 2000), a few years after the C. elegans genome was initially released. The genome contains 3 autosomal chromosomes (numbered 2-4), and one X chromosome. Each drosophila gene contains on average 4 exons of approx. 750bp a piece. Intron size is highly variable and can range from 40bp to more than 70kb. Introns and exons are both predicted to occupy around 20Mb of sequence.

3.5 Homo sapiens

Homo sapiens

Figure 3.4

Sequencing of the human genome was first formally proposed in 1985, but at the time the idea was met with mixed reactions in the scientific community. Then in1990 the Human Genome Project (HGP), under the direction by the N.I.H. and the Dept. of Energy, launched a 15-year, $3 billion plan for sequencing the complete human genome. Their progress was slow however and the HGP did not appear to be on pace to finish by the projected date in 2005. Half way through their planned time period, in early 1998, the HGP had sequenced less than 5% of the entire genome.

Then, in the same year that the HGP was reevaluating its progress, Celera, headed by Craig Venter, announced its intention to sequence the entire human genome over a three year period. After cutting their teeth on the Drosophila genome (which was done in collaboration with Gerald Rubin and the Berkley Drosophila Genome Project), Celera initiated the whole-genome shotgun sequencing of the human genome on September 8th, 1999. Less than a year later, on 17 June 2000, the first draft of the genome was completed. Today 99.9% of the human genome is 'finished', meaning less than 1 bp error per 10,000 base pairs.

The method Celera used, termed shotgun sequencing, is conceptually straightforward, but requires large amounts of computer processing power to complete. The protocol (in great oversimplification) is as follows: 1) cut up the genomic DNA into small pieces of known and regular size, 2) clone the pieces of genomic DNA into plasmids for purification and amplification purposes, 3) randomly sequence the DNA fragments from the plasmids while screening the results for contamination, 4) and then load the whole sequenced mess into the computer and let the computer sort it all out. The computer essentially plays a giant matching game building up larger and larger overlapping sequences until the whole genome is finally laid out in entirety. The process is, of course, not nearly this simple. One major complication worth mentioning is that the human genome is particularly replete with repeat sequences that could easily create numerous misleading matches. Computing the set of all overlaps required approx. 10,000 CPU hours on a suite of four-processor Alpha SMPs with 4 gigabytes of RAM (4-5 days in elapsed time using 40 such machines).

Celera's surprising and controversial success was due to several factors. First off, Celera was able to build upon the knowledge that previous sequencing efforts had gained through years of research and experience, including the Human Genome Project and The Institute for Genome Research (TIGR). Second, Celera's sequencing facilities were unparalleled in their sheer size. Celera's sequencing facilities had 50x the sequencing capacity of TIGR. Finally, because the results of the HGP were public, Celera was able to use their data to help align their shotgun sequences in the whole genome.

The human genome is 2.91-billion base pairs in length. Celera estimated that approx. 26,383 genes exist in the human genome, but this number has been a source of continued controversy with other estimates reaching as high as 150,000 genes (which is almost certainly much too high). Of the estimate 26,383 genes, 42% have an unknown function. The average number of exons in the predicted genes range between 4-5 and the typical exon length is around 100-300 base pairs. The average size of a human gene is around 27,000bp, with typical ranges between 20,000 and 50,000bp. A quick calculation will demonstrate that human genes are mostly intronic in composition. The average intron can be thousands of base pairs in size and can be as large as tens of thousands of base pairs (compare this to the typical exon with a paltry size of ∼200bp). Coding regions in the human genome are estimated to account for only around 3% of the total DNA sequence, intronic sequences contribute ∼30%, and intergenic regions ∼67%.

The expansion of non-coding DNA in humans is particularly striking when compared to other metazoan eukaryotes. For example, the human genome is 30x larger than the C. elegans and the Drosophila genome, but has only ∼2-3x as many genes. Furthermore, human genes are 10x larger than fly and worm genes, but the vast majority of this increase in size is due to intronic expansion; their exons are essentially the same size. Repeat sequences are another very prominent feature of the human genome. 35% of the entire human genome (including coding regions) is classified as repetitive, which is quite high already, but if we examine non-coding regions the proportion of repetitive DNA climbs to 46%. Compare these numbers to Arabidopsis which has a relatively low percentage of repeat sequences in the genome, 10%. But you should also keep in mind that Vivia faba, or the humble broadbean, is composed of upwards of 80% repetitive DNA.

Another important feature of the human (and other mammalian) genomes is CpG islands. A CpG island is a region of DNA that has a higher relative proportion of CpG dinucleotides when compared to the entire genome. This increased CpG density is significant because these regions tend to be unmethylated and therefore are believed to promote the initiation of transcription. This belief is drawn mainly from two observations: 1) most of the housekeeping genes (which are constitutively expressed genes) have CpG islands at the 5' end of the transcript, and 2) CpG island methylation is known to correlate with gene inactivation during gene imprinting and tissue specific gene expression.

3.6 Mus musculus

M. musculus

Figure 3.5

The mouse genome was sequenced by the Mouse Genome Sequencing Consortium in 2002 (Nature, Dec. 2002). Like the human genome, the mouse genome is large, 2.5Gb, only 14% smaller than the human genome. Gene prediction techniques estimate that there are 30,000 protein-coding genes in the genome. Approx. 99% of mouse genes have a direct, assignable human homologue. These genes are distributed among 19 autosomal chromosomes and one X chromosome. The mouse genome contains fewer CpG islands than the human genome (15,550 compared with 33,000) and, like the human genome, a large proportion of the mouse genome is composed of lowcomplexity repeat sequences. Sequencing the mouse genome was particularly important for a couple of reasons: the mouse is a ubiquitous as a research model, and for use as a comparative tool against the human genome.

Chapter 4

The Challenge[1]

Hello!

[1] This content is available online at <http://cnx.org/content/m11321/1.2/>.

Chapter 5

The Methods[1]

[1] This content is available online at <http://cnx.org/content/m11322/1.1/>.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Genefinding*
Edited by: Andrew Hughes
URL: http://cnx.org/content/col10205/1.1/

Module: "Introduction to Genefinding"
By: Andrew Hughes
URL: http://cnx.org/content/m11315/1.8/
Pages: 1-7

Module: "OLD * Genetics Background"
Used here as: "Genetics Background"
By: Andrew Hughes
URL: http://cnx.org/content/m11320/1.2/
Pages: 9-13

Module: "Genomic Data Sets"
By: Andrew Hughes
URL: http://cnx.org/content/m11317/1.3/
Pages: 15-20

Module: "TODO : The Challenge"
Used here as: "The Challenge"
By: Andrew Hughes
URL: http://cnx.org/content/m11321/1.2/
Page: 21

Module: "The Methods"
By: Andrew Hughes
URL: http://cnx.org/content/m11322/1.1/
Page: 23

Genefinding

Genefinding - the basic structure of the course attempts to answer the following questions: 1) What is genefinding? 2) Why do we care? 3) Why is it difficult? What is the challenge? 4) What's being done currently? What are the current methods? 5) How are we doing? Emphasis is placed heavily on computational methods for genefinding with discussion as to how the compliment experimental gene finding methods.