
© Photodisc/Getty Images.

Chapter 1

Bioinformatics and Genomic Data:
Investigating a Complex Genetic Disease

Chapter Overview

This is a skill-development chapter: it does not address a specific bioinformatic algorithm. The goal of this chapter is to build skills in retrieving information from genomic databases; it is appropriate for both programming and nonprogramming courses but could be skipped if students are already familiar with genomic databases and genome browsing. For nonbiologists, the BioBackground box at the end of the chapter provides a basic tutorial on genes and genomes.

Biological problem: Genes associated with Parkinson disease

Bioinformatics skills: Searching genome databases and metadatabases, genome browsing

Bioinformatics software: Entrez interface to NCBI databases, UCSC genome browser

Programming skills: None

Understanding the Problem:
Parkinson Disease, A Complex Genetic Disorder

At least seven million people are currently living with Parkinson disease (PD), a severe, progressive, and incurable disorder of the central nervous system. Actor Michael J. Fox's candid discussion of his condition has helped to focus attention on this disease, which begins with shaking, stuttering, difficulty walking, or involuntary muscle movement and worsens over time. PD is a ***complex*** *or* ***multifactorial*** *disorder: It has a heritable component but cannot be attributed to any single gene. Like autism, type 2 diabetes, asthma, or obesity, it likely results from the interaction of multiple genes as well as environmental or developmental components, complicating research into causes and treatments. We know that PD symptoms result from the death of a population of brain cells (neurons) residing in an area called the substantia nigra that normally produce the neurotransmitter dopamine. However, the cause of these cells' demise is unclear, and whereas a small percentage of PD patients have clearly defined defects in one of several specific genes, most do not, leaving biomedical researchers with no specific target toward which therapy can be directed. A genetic component of the disease is suggested by the fact that close relatives of Parkinson patients are at higher risk, but there is no clear pattern of inheritance; understanding of the disease is further complicated by environmental risk factors, including tobacco smoke.*

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

Parkinson's disease is just one example of a biological problem whose solution will depend heavily on bioinformatics. The cause of many well-studied genetic diseases can be narrowed down relatively easily to the inheritance of a dysfunctional allele of a single gene from one (in the case of **dominant** genetic disorders such as Huntington disease or hypercholesterolemia) or both (for **recessive** disorders such as cystic fibrosis, sickle-cell anemia, phenylketonuria, and Tay-Sachs disease) parents. Identifying the genetic factors in PD, however, is a more difficult undertaking that requires computational tools to facilitate the analysis of large amounts of genomic data.

A working draft of the nucleotide sequence of the entire human genome was completed in 2001, adding nearly 3 billion nucleotides to publicly accessible databases such as GenBank. Today, thousands of different genomes from plants, animals, bacteria, archaea, fungi, viruses, and protists have been completely sequenced, and new data continue to be added rapidly as the cost of genome sequencing projects continues to fall (**Figure 1.1**). These genomic data are used not only to investigate diseases but are also used by research scientists in areas as diverse as molecular biology, physiology, evolutionary biology, immunology, and ecology and by doctors, pharmaceutical companies, public health officials, animal breeders, food scientists, sociologists, law enforcement agencies, and many others. Diagnosing genetic diseases, determining evolutionary relationships, understanding metabolic functions, designing new drugs, investigating forensic evidence, improving food supplies, tailoring medical treatments to individuals, and reversing environmental degradation are just a few of the numerous current and potential applications of sequence data.

Making sense of these hundreds of billions of base pairs of genetic information is a daunting task. Simply printing out the human genome sequence would require some 250,000 pages, double-sided and single-spaced. **Bioinformatics** is the new science at the interface of molecular biology and computer science that seeks to develop better ways to

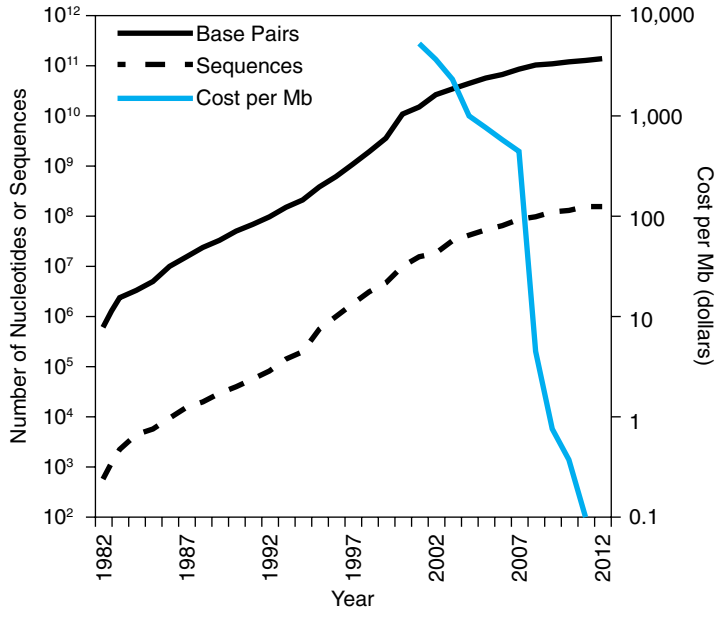

Figure 1.1 Nucleotide sequence data stored in the public GenBank database has grown exponentially for three decades, while the cost of large sequencing projects has declined dramatically. Data from: National Center for Biotechnology Information.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

explore, analyze, and understand this vast wealth of genomic data. It is a branch of **computational biology** (the application of mathematical methods and computer algorithms to biological problems) that focuses specifically on the storage, retrieval, manipulation, and analysis of DNA and protein sequences as well as the information on structure, expression, and so on that can be derived from them. Bioinformatic tools provide one avenue for better understanding complex disorders such as PD, increasing our ability to work toward improved treatments or cures. Our goal is to help students become familiar with the algorithms and software used to address key questions in bioinformatics through the use of existing Web-based tools and/or individual programming solutions to investigate genuine biological questions. In this chapter, we explore how databases that store and link genomic information can allow us to investigate and better understand complex biological problems like Parkinson's disease.

Bioinformatics Solutions: Databases and Data Mining

Although most programs and algorithms discussed in this text involve ways that new data are generated, analysis—often referred to as "**data mining**"—of existing information in genomic databases can also yield valuable new insights. The nucleotide sequence of the entire human genome has been determined and is publicly accessible, but we have not yet unambiguously identified all of the genes within the genome (see the chapters on gene prediction), let alone determined their functions (see the chapter on protein alignment). However, in addition to sequence information, we have access to data about phenotypes, expression, intron/exon prediction, transcription factor binding, and more. "Mining" databases that bring together these different types of information can allow us to make new discoveries about known sequences, especially when the power of the bioinformatic data is combined with experimental results.

Except for the small proportion of PD patients whose disease clearly results from mutation in one of several specific genes, the genetic contribution to PD has generally been thought of as relatively small: Only about 15% of PD patients have a parent, sibling, or child with PD. However, a number of regions of the human genome have been correlated with PD through experiments such as **genome-wide association studies** (**GWAS**) in which a large number of genetic differences known to occur among individuals (often identified through genome sequencing) are examined to identify possible links to a specific genetic disease. A 2011 study by Do et al. (see References and Supplemental Reading at the end of this chapter) identified several new regions of the genome associated with PD and suggested that genetic changes contribute to a much larger proportion of both early- and late-onset cases of PD than was previously believed.

GWAS, however, can only identify genome *regions* that may be associated with a disease, not specific genes or specific mutations. Further analysis is necessary to determine what gene or genes are encoded in the identified regions, which of those genes are likely to be involved in the disease of interest, and what specific mutations may account for the disease phenotypes; this is where bioinformatics comes in. In this chapter's projects, we use genomic databases and metadatabases and the University of California Santa Cruz (UCSC) genome browser to examine the PD-associated genome regions identified by Do et al. and see how mining these databases using bioinformatic techniques can enable us to formulate hypotheses about which specific genes in the identified regions could be involved in PD. Such hypotheses can guide future research into the causes, prevention, and treatment of Parkinson's disease.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

BioConcept Questions

BioConcept questions test your understanding of the biological ideas needed for each chapter. If you find that you need some help with these concepts, the BioBackground boxes are intended to provide a working knowledge sufficient to complete the chapter projects; these can be skipped if you are already clear on the concepts tested here.

1. In thinking about genetic diseases, it is common to refer to a "disease gene," such as the cystic fibrosis gene or the Tay-Sachs gene. A geneticist, however, would insist that we should refer to a "disease *allele*" instead. Why is this seemingly subtle difference an important one?
2. A recessive genetic trait (such as cystic fibrosis, sickle-cell disease, or simply short eyelashes) is only shown phenotypically if an individual inherits the recessive allele from *both* parents. If a mutation results in an allele that does not encode a functional protein, it is generally true that the allele is recessive. Can you explain why nonfunctional alleles are most often recessive and not dominant?
3. If every cell in the human body has the same DNA, how can they have such different structures and functions?
4. Briefly outline the pathway of gene expression, starting with DNA and ending with the protein product of a gene.
5. Which is longer, a gene's transcript or its coding sequence? Why?
6. When you see DNA represented as a string of letters (such as AACGATCC . . .), what do those letters represent? Why isn't it necessary to write out both strands?
7. Does the sequence `CAGCCUCCGA` represent a DNA sequence or an RNA sequence? How do you know?
8. When you see a protein represented as a string of letters (such as `MFRVAMP` . . .), what do those letters represent?

Understanding Genomic Databases

Learning Tools

From the text's website, you can download an HTML file, `DNASidebar.htm`, containing links to all Web databases, tools, and other sites mentioned in this text. Firefox users can load this list into the sidebar to provide a "bookmark" list that is always visible while working on other things. To do this, save the file to your computer and then open it in Firefox. Create a bookmark to the page and then edit the bookmark properties and check `Load this bookmark in the sidebar`. When you access this bookmark, the link list appears as a sidebar alongside whatever page you are viewing. Throughout the text, the link icon (the wrench shown to the left) denotes a website whose URL can be found in the list in `DNASidebar.htm` or at the *Exploring Bioinformatics* website. Websites that correspond to the link icon will be in bold blue type.

Primary Databases

Since molecular geneticists began to acquire significant DNA sequence information in the late 1970s, they have emphasized the importance of making these data widely available. The vast majority of all known gene and genome sequences have been deposited in databases accessible to scientists worldwide—and to the general public—via the Internet. The three major databases for DNA sequence information (each of which mirrors the others) are (1) GenBank, maintained by the National Center for Biotechnology Information (NCBI),

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

a unit of the National Library of Medicine, which in turn is a branch of the U.S. National Institutes of Health; (2) the European Molecular Biology Laboratory (EMBL) database; and (3) DNA Data Bank Japan, maintained by the National Institute of Genetics of Japan. On the protein side, the UniProt (Universal Protein Resource) database is considered the most comprehensive; it combines data formerly maintained in three distinct repositories. These are called **primary databases**, because this is where "raw" nucleotide or amino acid sequence information is deposited. They are also **annotated databases**, because database records contain additional information about the sequences, such as the locations of protein-coding regions, introns and exons or other genetic features, as well as references to the scientific literature. Additional primary databases have also been created to store specific kinds of data such as the results of gene **expression** experiments, known **polymorphisms**, and so on. Table 1.1 shows a list (by no means comprehensive) of some useful primary databases.

When we retrieve a DNA or protein sequence from a database, typically we see it laid out in some clear, readable format, usually within a Web browser. Consider, however, the problem of how the raw data should be stored. It is in fact not very practical to store formatted data, because this reduces the likelihood that a different application can effectively access it or that it can be readily repurposed as needs and software change. Like any database, then, genomic databases are divided into records (sequences) and then into fields. There are fields for the raw sequence itself; for the locations of features such as coding sequences, promoters, or introns; and for annotations such as references or additional information. Formatting is left up to the software that retrieves the sequence. Ideally, it should be easy for another database to retrieve a desired piece of information and connect it to related information stored elsewhere to generate a metadatabase.

Metadatabases

Along with the enormous growth of sequence information has come a need for additional resources that assist researchers in finding and retrieving data and, increasingly, finding the interconnections among the various kinds of stored data. **Secondary databases** or **metadatabases** (as the term is used by bioinformaticians) select and combine data from other databases (Figure 1.2). For example, **NCBI's Gene database** pulls together the DNA sequence, the protein sequence, references and information on expression, alleles and phenotypes, genomic location, and much more for genes that have been well studied. The **OMIM (Online Mendelian Inheritance in Man) database** brings together a wealth of information about all the known human genetic diseases and the genes that contribute to them. Or, one might choose the **KEGG Pathways database** to focus on metabolic pathways and the genes that encode metabolic proteins. Table 1.1 also lists a few useful metadatabases.

Database Searching

Searching a database requires some form of user interface. Today, this is most often a Web-based interface, with the data stored on a remote server. The search interface is not part of the database itself, and multiple Web-based interfaces can be found that all search the same underlying database. It is of course impossible to comprehensively describe these search interfaces here, but Box 1.1 provides syntax information for the Entrez search interface. Entrez is the common interface for all databases maintained by NCBI—databases that are heavily used in this text and by actual researchers.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

Table 1.1 Summary of some useful databases and resources.

Resource and Location	Description
NCBI Databases	
Nucleotide www.ncbi.nlm.nih.gov/nuccore	Main interface to annotated nucleotide sequences in GenBank and other major repositories
Protein www.ncbi.nlm.nih.gov/protein	Main interface to annotated protein sequences and translated DNA sequences from GenBank and other major repositories
Gene www.ncbi.nlm.nih.gov/gene	Metadatabase compiling information on genes from well-annotated genomes, including maps, sequences, functions, expression, etc.
OMIM www.ncbi.nlm.nih.gov/omim	Database of human diseases and genes associated with human disease; includes entries for both the diseases and the genes
Map Viewer www.ncbi.nlm.nih.gov/mapview	Genome browser: graphical view of genes and sequences in the context of chromosome, phenotype, marker, single-nucleotide polymorphism (SNP) and other maps
Gene Expression Omnibus (GEO) www.ncbi.nlm.nih.gov/geo	Composite of results from a large number of gene expression experiments
HomoloGene www.ncbi.nlm.nih.gov/homologene	Metadatabase focused on showing conservation of genes and identifying their orthologs
dbEST www.ncbi.nlm.nih.gov/dbEST	Database of expressed sequence tags: sequences of cDNAs representing fragments of genes expressed in some tissue or condition
dbSNP www.ncbi.nlm.nih.gov/dbSNP	Database of known SNPs
Other Databases	
KEGG www.genome.jp/kegg	Metadatabase focused on protein structure and function
Human Gene Mutation Database www.hgmd.cf.ac.uk/ac	Database of known human mutations (registration required)
PDB www.pdb.org/pdb	Database of protein structures and nucleic acid secondary structures
KEGG Pathways www.genome.jp/kegg/pathway.html	Database of metabolic pathways linked to known genes and proteins
UCSC Genome Browser genome.ucsc.edu	Genome browser: graphical view of genome sequences, with tools for mapping sequences to the genome, visualizing expression, etc.
STRING string.embl.de	Database of protein interactions
Pfam swww.sanger.ac.uk/resources/databases/pfam	Database of protein families and domains
Wormbase www.wormbase.org	*Caenorhabditis elegans* genome and associated resources
Flybase www.flybase.org	*Drosophila melanogaster* genome and associated resources
SGD www.yeastgenome.org	*Saccharomyces cerevisiae* genome and associated resources
Colibri genolist.pasteur.fr/Colibri	*Escherichia coli* genome and associated resources

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

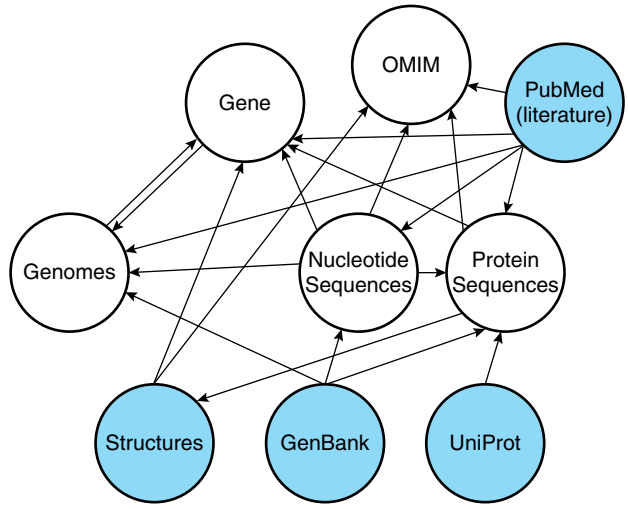

Figure 1.2 Example showing how primary genomic databases (filled circles) and metadatabases (open circles) might be interrelated.

Genome Browsers

A genome browser acts like a metadatabase in that it brings together information from many genomic databases, but it does so in a graphical form (**Figure 1.3**). A computer scientist might think of a genome browser as a graphical user interface (GUI) for genomic databases. Typically, a genome browser shows a graphical representation of the chromosomal position of a specified gene or genome segment. This view can be zoomed out to show more

Box 1.1 Syntax for the Entrez Search Interface to NCBI Databases

- **AND, OR,** and **NOT** must be capitalized.
- **Quotes** can be used to search for "exact phrases"; however, use quotes with caution, because the phrase will only be found if it occurs in the database's phrase list; the complete text of database entries is not searched.
- **Parentheses** can be used for grouping: `(CFTR AND complete) NOT human`.
- The asterisk is the wildcard character, representing any characters; therefore, `cys*` would find cystic fibrosis but also cysteine.
- The search can be limited to a particular database field using square brackets. For example, `smith [Author]` will limit the search to records with authors named Smith. Other useful fields are [Organism], [Title], [Text Word], and [Gene Name].
- Searching for a name followed by initials has the same effect as limiting to the author field: `smith j` or `smith je` would search for the author J. Smith or J. E. Smith, respectively.
- Click `Limits` on a search result page to include or exclude sequences by date or by other criteria appropriate to the particular database; the `Filter` option on the results page is another way to limit search results.
- Click `Advanced` to build a query using drop-down field lists and other tools or to run an additional query on a previous search result.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

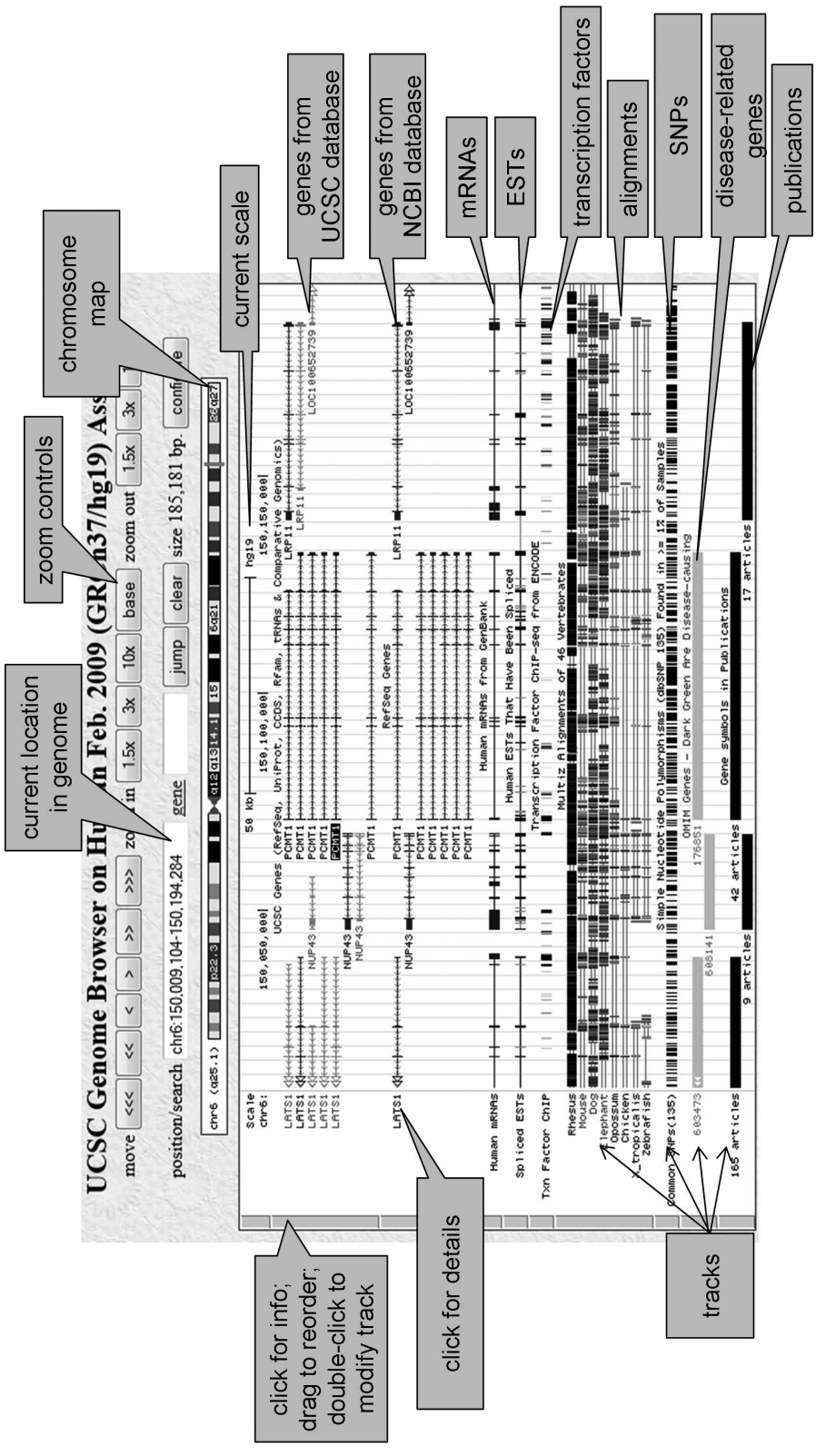

Figure 1.3 Sample display of a human chromosome region from the UCSC genome browser. Screenshot from UCSC Genome Browser: Kent et al., Genome Res. 12:996 (2002).

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

of the chromosome or zoomed in to show particular regions of a gene or even the actual DNA sequence. There are then various **tracks** that can be shown, hidden, or in some cases modified by the user. By default, one or more tracks representing genes are usually shown; "genes" in this case are typically defined as transcribed regions, so this track may show multiple known or hypothesized transcripts, and several different gene tracks can represent different database sources or types of evidence for transcription. Introns and exons are also represented in this view (if you are unfamiliar with concepts such as transcripts, introns, or coding sequences, see BioBackground at the end of this chapter), and each gene can be clicked to view more detailed descriptions, lengths, and references or to retrieve sequences or link to additional databases.

Additional tracks show binding sites for transcription factors, expression in specific tissues, the locations of known genetic variations (mutations or polymorphisms), methylation sites, repeated sequences, and comparisons with the genomes of other organisms. Users can even add custom tracks showing their own data. Genome browsers usually also have integrated tools for functions such as aligning a gene of interest with the genome or with the genomes of related organisms or predicting the sizes of **polymerase chain reaction** (**PCR**) products. Although all these data can also be accessed by other means, genome browsers have become popular due to the vast wealth of information consolidated in one location. Currently, the most-used genome browser is the **UCSC genome browser** maintained by the Genome Bioinformatics group at the University of California Santa Cruz. NCBI has its own genome browser, **Map Viewer**, as does EMBL, the **ENSEMBL genome browser**; many other genome browsers can be found online, some that are generally useful and some with specialized functions.

Test Your Understanding

1. Classify each of the following databases as primary databases or metadatabases.
   - **a.** GenBank
   - **b.** Gene
   - **c.** Protein Data Bank, a database in which structural biochemists deposit newly determined, three-dimensional structures of proteins
   - **d.** Colibri, a database providing information about each gene within the genome of the bacterium *Escherichia coli*
   - **e.** Stanford Microarray Database, a repository of the results of microarray experiments
   - **f.** TreeBASE, a database containing phylogenetic trees constructed based on nucleotide or protein sequence data
2. Using a Web search engine, identify two primary databases and two metadatabases not mentioned in this chapter and list the main goals or functions of each.
3. Although most sequence information has been deposited into one of the public databases, for-profit companies sometimes sequence genes or genomes and keep the information private, at least for a period of time. Discuss the value of public sequence data versus the need for industry to control access to information that may affect their ability to produce their products.
4. Navigate to the **UCSC genome browser** and choose any chromosome region to display. Identify two tracks not shown by default and list their functions.
5. In the sample genome browser display in Figure 1.3, several different bars are labeled as representing the same gene. These bars are sometimes, but not always, the same length. If these bars all represent one gene, why do you believe there are different bars, and why are they not identical?

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

CHAPTER PROJECT:

Genomic Regions Associated with Parkinson Disease

Do et al. (see References and Supplemental Reading) used a GWAS to look for genetic determinants of PD. GWASs, however, do not identify specific genes but rather use markers that permit the identification of genome regions correlated with the trait of interest. The researchers must then begin identifying candidate genes in the regions found, using bioinformatics to generate hypotheses and then testing them with laboratory research. In this project, we use genome databases and the UCSC genome browser to develop testable hypotheses about genes that might be involved in PD based on the regions identified by Do et al.

Learning Objectives

- Understand the kinds of information stored in genomic databases and metadatabases
- Develop skill in using various interfaces to find and retrieve genomic information
- Gain familiarity with the use of a genome browser to examine genomic regions and identify genes and other features
- Appreciate the value of bioinformatic data mining to develop hypotheses for further research

Suggestions for Using the Project

This chapter includes a Web Exploration Project but no Guided Programming Project. Its intent is to develop skills in the use of Web-based genomic databases and tools that will be needed in future chapters. This project is recommended both for courses that require programming skills and those that do not because of the familiarity the students will gain with how genomic information is stored and used and how computers represent and manipulate DNA and protein sequence information. Instructors should feel free to develop exercises of their own that use these same skills. An optional On-Your-Own Project is included for instructors who would like their students to practice their skills further.

Web Exploration: Data Mining in the Genome Databases

A GWAS requires a method to determine many individuals' **genotypes** for a large number of sites in the genome where genetic variation is known to occur. This is often done by means of a DNA microarray using allele-specific oligonucleotides: Short, single-stranded DNA segments with sequences matching known polymorphic sites are allowed to base pair with DNA from a particular individual under conditions where even a one-base mismatch would prevent binding. Do et al. genotyped 3,426 PD patients and 29,624 healthy control individuals for 522,782 known **simple nucleotide polymorphisms** (**SNPs**), which are places in the genome where variation occurs among individuals in the form of a single nucleotide change or a one- or few-base insertion or deletion. They identified 11 SNP sites where one allele was correlated with PD with a statistically significant frequency.

Part I: Genome Browsing to Identify Possible Disease-Associated Genes

Exploring Bioinformatics on the Web

Find quick links to Web-based tools (updated as URLs change) at the *Exploring Bioinformatics* website.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

Genome browsing for SNPs. Known SNPs in the human genome are deposited in the primary genomic database dbSNP; each has a unique **accession number** that identifies it. One SNP identified by the Do et al. GWAS is rs11868035; we start by investigating the genomic neighborhood of this SNP, using the **UCSC genome browser**. Note the following:

- If you use a Web search to find the genome browser, you may land on the UCSC Genome Bioinformatics main page; if so, find the genome browser in the list on the left side.
- The genome browser saves your preferences; if you have used this program before (or someone else has used it on the same computer), click the `Click here to reset` link before continuing to be sure your experience matches the discussion that follows.

You can use the UCSC genome browser to browse a variety of genomes; by default, the search parameters should be set to the latest version ("assembly") of the human genome, which is what we want in this case. You should see an input box where you can enter a specific position in the genome or a search term: gene names, key words, and accession codes all work here. Using rs11868035 as your search term, search the genome for the SNP identified by Do et al.

From the results page, you can see various data sets that include this SNP. Of specific interest to us is the listing for the NHGRI Catalog of Published Genome-Wide Association Studies, because this means the SNP has been identified as a point of interest in at least one GWAS. Because the SNP occurs at a specific point within the genome, any of these links will show you that genome region in the browser; however, different links will turn on different tracks that may provide different kinds of information. (You may wish to review Figure 1.3 to familiarize yourself with the track interface.) For our purposes, click the NHGRI link to find the SNP in the genome.

Examine the resulting browser view and get a feel for the kinds of information displayed. Notice from the scale bar (see Figure 1.3) that you are zoomed in to see a very small region of the genome: About how many nucleotides are displayed in this view? Just below the scale bar, you should see a track labeled NHGRI Catalog of Published Genome-Wide Association Studies; this track was turned on because of the link you selected and shows that the genome view is centered on the rs11868035 SNP. The point where the SNP occurs is shown by a bar, and its accession number is highlighted. Just above this display, you should also see an **ideogram**, or schematic drawing of the chromosome where this SNP occurs; the dark and light regions represent bands that are visible when real chromosomes are stained. Bands along the p (short) and q (long) arms of the chromosome are numbered, so that a chromosome position can be described by notation like 6q25.1. What is the chromosome position for the rs11868035 SNP?

Near the bottom of the track display, you should see a track labeled Simple Nucleotide Polymorphisms. Here, you should see bars for other SNPs in this region. To see their labels, right-click the track title and choose `full`. The rs11868035 SNP should again be highlighted. If you click its accession number in this track, you will jump to a page of information from the dbSNP database where you can see what nucleotides have been observed in human alleles at this position: In this case, the SNP is a base substitution, and either A or G can occur at this position, giving two known alleles. You can also click the SNP's accession number in the NHGRI Catalog track; you should see that it has been identified in the Do et al. study, as expected. Has it also been identified by other GWASs?

Returning to the track display, try zooming out by clicking on the `10× zoom out` button. As you zoom out, you will see more SNPs and begin to see the boundaries of genes (change the display of the SNP track back to `dense` to keep it from getting excessively long). At the top of the track display are three tracks that show different views of genes in this

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

chromosome region: UCSC Genes, RefSeq Genes, and Human mRNAs. The labels on the left give the official gene symbol ("name") for each gene. Thick lines represent segments of coding sequence, called **exons**, whereas thin lines indicate **introns**, which are sequences that interrupt the coding sequence and are spliced out after transcription. In the UCSC Genes track, introns are shown with arrows to show the direction of transcription and transcribed, but untranslated regions (UTRs) flanking coding sequences are shown by medium-width lines. Does the rs11868035 SNP occur within a gene? If so, what is the name of the gene?

Now, zoom in far enough to read the actual nucleotide sequence by clicking on the `base` button. The display should still be focused on your highlighted SNP; if you have changed the focus (such as by clicking on a gene), you can always zoom out, find your SNP in the NHGRI or SNP track, right-click it, and choose `Zoom to . . .` to recenter. With the display zoomed in to this level, you should be able to see the position of the SNP within the gene: Notice in this case (as shown by the Human mRNA track) that the SNP is within an intron, not within the coding sequence itself. Mutations within introns—especially if they occur near the boundary between an intron and exon—can affect the expression of a gene by impeding the splicing process. However, this could also indicate that the SNP identified in the GWAS occurs within a chromosome region important in PD but not within the actual *gene* correlated with the disease. We need additional evidence before we decide to focus on this particular gene.

Evidence for possible disease involvement. One line of evidence that a particular genomic region is important is **conservation**, which is the maintenance of the region in the genome over evolutionary time. A track labeled Multiz Alignments should be visible by default; this track shows an **alignment** between the human genome region you are browsing and the genomes of other sequenced vertebrates. Is the gene in which the rs11868035 SNP is found conserved? Is the specific *sequence* where this SNP occurs conserved? What animals might be appropriate model systems in which to conduct further research into the role of this gene in PD?

Although the Do et al. GWAS identified the rs11868035 SNP, this was just a marker in the study, and in fact any genes in its neighborhood could be responsible for the correlation between this region and PD. To analyze additional evidence, we need to turn on some tracks that are not displayed by default. First, zoom out so you can see a larger genome region surrounding the SNP of interest: Get a view showing approximately 1,000 kilobases (1,000 kb, or 1,000,000 base pairs) of the genome. This should bring the entire length of several genes into view, still centered around your original SNP. Now, scroll down below the track display, where you should see a number of drop-down boxes allowing you to turn tracks on and off. Under Phenotype and Disease Associations, find a drop-down list labeled `GAD View` and choose `pack` from the list. Click `refresh` to display the result of your change. This track will show previously described associations of genes with complex disorders. Similarly, turn on the track labeled `OMIM Genes`, showing genes listed in the OMIM database associating human genes with phenotypes, including disease phenotypes. Also, under Expression, turn on the `GNF Atlas 2` track, which summarizes experiments testing expression of the genes in various tissues. Finally, under Mapping and Sequencing Tracks, turn on the `Publications` track if it is not already on (you may want to choose the "dense" option for this one). Refresh the display.

Now, what kind of evidence would suggest the possible involvement of a particular gene in PD? Because PD is a disorder of the central nervous system, genes expressed in the brain or in nervous system tissue are good candidates, as well as any genes previously associated with some type of neurological disorder. Genes shown in red in the GAD View track have been associated with some kind of disease; click them to see if any seem relevant to PD. The OMIM

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

Genes track uses a color scale to show how clearly a gene has been associated with a disorder or phenotype; click one of the green or gray bars for a page that describes this scale. Then click the OMIM entry number link to see a summary of information from the OMIM database. OMIM catalogs both genes and disorders, so from the summary page, you can click either the gene link or the disorder link to retrieve entries from OMIM. By mining the data others have already found, can you strengthen the case for involvement in PD for any of these genes?

In the GNF Expression Atlas track, red bars represent genes that were strongly expressed in the indicated tissues (see labels at left): the brighter red, the more expression. Black represents a gene that is neither over- nor underexpressed, and green represents a gene that is underexpressed in the given tissue. A gene that is expressed in brain or nervous system tissue would certainly suggest a possible link to PD, but so might a gene normally underexpressed in these tissues turned *on* as the result of a mutation. Is there evidence that any of these genes are expressed in appropriate tissues?

Finally, click on each gene's bar in the Gene Symbols in Publications track to see a summary of published papers that refer to it. If you are in dense view, you will need to click twice, once to expand the gene list and then again to choose a particular gene. Is there any experimental evidence to suggest possible involvement of this gene in a neurological disorder? Taking together various pieces of evidence such as this, Do et al. concluded that the genes in this region most likely to be involved in PD were *SREBF1* and *RAI1*. Do you agree with their analysis? What evidence did you find that would support this conclusion?

Web Exploration Questions

You have seen how the UCSC genome browser can be used to expand on a PD-associated SNP found by a GWAS, leading to the identification of one or more candidate genes for further study. Now, use the skills you have learned to investigate a second SNP reported by Do et al.: rs34637584. Answer the following questions about that SNP:

1. On which human chromosome is this SNP located and at what position? (Use the conventional chromosome position notation described earlier.)
2. Does rs34637584 occur within a gene or between genes? If it occurs within a gene, is it within an exon or an intron?
3. Describe the alleles that occur through variation at this site.
4. List the genes found within approximately 500 kb on either side of the SNP.
5. Has this site or any of the nearby genes been associated with PD previously?
6. What evidence did you find to support the identification of one or more of the genes in this region as a candidate for a PD-associated gene?

Part II: Retrieving Sequences and Examining Genes in Detail

As you have seen, the UCSC genome browser is a powerful tool for genome analysis, bringing together a vast wealth of data that can be mined to answer many different kinds of questions. Similar analysis could be done by using **NCBI's Map Viewer**, the **ENSEMBL genome browser**, or any of a variety of related tools. In fact, without leaving the genome browser, you could retrieve the DNA or protein sequence of a gene you are interested in and go on to the kinds of analysis described next. For the purposes of this exercise, however, we will retrieve sequences by searching GenBank (a primary repository of nucleotide sequences) directly, so that we can learn to use the NCBI Entrez interface. We will then learn more about them through the use of metadatabases.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

You found that the rs11868035 SNP was located within the gene *SREBF1*. Suppose you now want to download the sequence of this gene and/or the protein it encodes, perhaps to clone the gene for further experimentation or to perform a more detailed bioinformatic analysis such as alignment, structure prediction, or phylogenetic analysis. To accomplish this, go to the **NCBI Nucleotide database**. All NCBI databases can be searched using an interface called Entrez; once you find the NCBI site, start with a simple search by typing `SREBF1` into the search box at the top of the page and choose `Nucleotide` as the database to search using the drop-down box.

You will quickly see that this search may have been a little too general. Although you may see useful sequences in the result list, you will also recognize that you have more than 100 total results, only some of which represent human genes. A narrower search might be easier to interpret; one way to limit the search to human genes would be to use `SREBF1 AND homo sapiens` as your search term; note the AND must be capitalized (see search tips presented in Box 1.1). Another way to accomplish this is to look at the list of matched organisms on the right side of the results page (a variety of tools and links are listed here) and click on the link next to *Homo sapiens*. You should also see a list of "filters" that allow you to pare down your results further: Clicking on the `RefSeq` link, for example, will leave only those that are also found within NCBI's Reference Sequence database, which is a subset of GenBank containing nonredundant, well-characterized sequences. Note also that some of the results listed are not actually for *SREBF1* but for some nearby gene (whose record probably mentions *SREBF1* somewhere in its text). To eliminate these, you could use `SREBF1 [Title]` or `SREBF1 [Gene Name]` as search terms to require a match to a particular field.

Even with these limits, you will still see more than one result. Some matched records include "mRNA" in their titles, indicating they are sequences of cDNAs—mRNAs copied to DNA by the enzyme reverse transcriptase and then sequenced—rather than sequences of genomic DNA. For our purposes, look for an entry that includes "RefSeqGene" in its title; it should have the accession number NG_029029.1. Click on this sequence.

By default, you will see the complete GenBank record for this gene, including description, references, and a listing of a variety of features. The actual nucleotide sequence is at the bottom of the record: Are you surprised by its length? To see only the nucleotide sequence, click on the link at the top labeled `FASTA`; this will show the sequence in a conventional format called FASTA (introduced in a popular alignment program of the same name) in which the sequence (with no spaces or numbers) is preceded by a one-line descriptive comment marked by the > symbol (**Figure 1.4**).

Return to the GenBank display and consider the features list. This list shows the location of important features within the sequence, like the coding sequence (designated CDS in a

```
>gi | 28302128 | ref | NM_000518.4 | Homo sapiens hemoglobin, beta (HBB), mRNA
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGGTCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

Figure 1.4 The gene encoding the beta chain of human hemoglobin shown in FASTA format.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

GenBank entry), the mRNA, and the various exons. Clicking on the links associated with these features alters the sequence display to show only the desired feature. Or, the locations of the features within the sequence can be readily visualized by choosing `Highlight Sequence Features` from the list of links on the right side of the page and then choosing the desired feature from the drop-down box that appears at the bottom of the page. Try highlighting the gene, then the mRNA, then the CDS, and compare the results.

The list on the right also links to other resources with additional information about this gene. You will recognize, for example, the link to OMIM as a database where you could learn about disease or phenotype associations of this gene. PubMed is a database of scientific publications where you could look up what has been published. For each protein-coding gene indexed in Nucleotide, there is a corresponding entry in Protein for the amino acid sequence that can be linked from this list. Another valuable choice from this list is the Gene metadatabase, where NCBI has compiled details about well-studied genes from a variety of sources. Here, you will see a graphical display of the gene in its genomic context and a mini genome browser showing a close-up view of the gene with its introns and exons. Many kinds of information are available on this page or through links or roll-overs, some of which should be familiar to you from the previous section of this project.

Web Exploration Questions

Mining the resources in the GenBank entry and the Gene metadatabase provides valuable information to a researcher seeking to better understand this gene or its physiological roles and should enable you to answer the questions below.

7. How long is the entire *SREBF1* gene (in bp or kb)?
8. When you click on the CDS link in a GenBank entry, only short segments of the gene sequence are highlighted. What do these highlighted segments represent?
9. How long is the spliced mRNA for *SREBF1*? What fraction of the gene is thrown away in the form of spliced-out introns?
10. What accounts for the difference between the sequence segments that are highlighted when you click on the `mRNA` link versus when you click the `CDS` link?
11. How long is the SREBF1 protein in amino acids? What are the first 10 amino acids in the protein sequence?
12. What is known about the function of this gene?
13. Besides its hypothetical association with PD, what are two other known connections of *SREBF1* to disease?

On-Your-Own Project: Clues to a Genetic Disease

Now that you have seen some genomic databases and worked with sequence information in a variety of ways, you should be able to apply these skills to a new project. Choose a genetic disease of interest to you (your instructor may require that you choose a single-gene disease or a complex disease), identify a gene associated with this disease, and summarize in one or two typed pages basic information about the gene, the protein it encodes, its genomic location, its expression, its known function(s), and its association with the disease you chose. OMIM would be a great place to start. If it is a gene that has been clearly identified as the specific cause of the disease, tell how the disease allele differs from the normal allele, if that is known. Document the sources of your information and give at least two references to published papers where your reader could learn more.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

BioBackground: Genomes, Genes, Alleles, and DNA

An organism's **genome** is the complete set of all its **genes**, each of which can in turn be thought of as the encoded instructions for synthesizing one protein. All members of a species have the same set of genes—genes encoding hemoglobin and digestive enzymes and hair proteins and eye lens proteins—every protein that any cell in the organism can possibly make. Every cell within the organism carries this same genome. Each gene is a segment of **DNA** (deoxyribonucleic acid), and the genes are joined together to make up a set of very long DNA molecules called **chromosomes** that reside within the **nucleus** of every cell.

Individuals vary because although they all have the same set of genes, they do not all have the same **alleles**. An allele is a specific form of a gene: If a gene is thought of as a sequence of the DNA nucleotides A, T, C, and G, then one individual might carry the allele of the hemoglobin gene that begins `ATGGTGCATCTGACTCCTGAGGAG` and another individual might carry an allele beginning `ATGGTGCATCTGACTCCTGTGGAG`. In both cases, this is the same gene, encoding the hemoglobin protein, but different alleles (ultimately arising from mutations) encode slightly different variants that may function differently. Genetic variations are also known as **polymorphisms**.

Actually, each individual—and each cell within that individual—carries *two* complete sets of genes: one inherited from the mother and one from the father. So, you might inherit an allele of the "dimple" gene that produces dimpled cheeks from your mother and an allele of the *same* gene from your father that does not. In this case, you would have dimpled cheeks: The dimpled allele is said to be **dominant** over the nondimpled allele—it only takes one dimpled allele to produce the dimpled **phenotype**, or observable characteristic. On the other hand, having round eyes is a **recessive** trait: It is necessary to inherit a round-eye allele of the eye-shape gene from *both* parents to show this phenotype; the almond-eye allele is dominant. Simple genetic diseases result from alleles that encode an abnormal version of a protein; these alleles can be recessive (as in the case of sickle-cell anemia, caused by a misfunctioning version of the hemoglobin protein: see **Figure 1.5**) or dominant. An individual randomly inherits one of the mother's two alleles and one of the father's two alleles; thus, if the parents' genetic makeup (**genotype**) can be determined, the laws of probability can be used to determine how likely it is that their children will show a particular trait.

The cells within one individual vary not because they carry different genes or alleles but because different types of cells *express* different genes. We can think of a gene as a segment of DNA

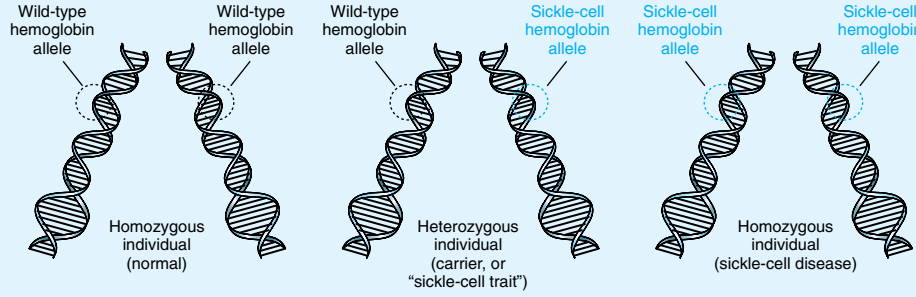

Figure 1.5 An individual has two alleles of each gene, either the same or different. The alleles inherited from the individual's parents determine his/her phenotype—in this case, whether s/he will have sickle-cell disease, normal hemoglobin, or carry the sickle-cell trait.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

that encodes a protein (although some genes encode RNAs that are never translated to proteins), and **gene expression** is the process of actually making that protein. This is a two-step process: **Transcription** (carried out by **RNA polymerase**) copies nucleic acid information for one gene from DNA to **messenger RNA** (**mRNA**), and **translation** (carried out by the **ribosome**) decodes that information to make the protein (Figure 1.6). This DNA → RNA → protein process is so fundamental to all of biology that Francis Crick dubbed it the "central dogma of molecular biology."

In bioinformatics, DNA, RNA, and proteins are all represented as simple strings of letters, making them easy to manipulate computationally. A closer look at their structures reveals why this works. DNA (Figure 1.7) is a nucleic acid molecule consisting of two very long chains of **nucleotides** or "**bases**" (an average human chromosome is more than 100 million nucleotides long). The four nucleotides, A, T, C, and G, can occur in any order, and it is the specific sequence of nucleotides that identifies a gene, protein binding site, or other functional feature of DNA. The two chains are held together by **base-pairing**: A always pairs with T and C always pairs with G. Base pairing means that it is only necessary to write out the nucleotide sequence of one DNA chain, such as `ATGGTGCATCTGACTCCTGAGGAG`, to represent a double-stranded DNA molecule that really looks like

```
ATGGTGCATCTGACTCCTGAGGAG
TACCACGTAGACTGAGGACTCCTC
```

. Figure 1.8 represents the same process of gene expression as Figure 1.6, but now a string of letters representing nucleotides takes the place of the double-stranded DNA molecule.

Where a gene occurs within the DNA sequence, a nucleotide sequence known as a **promoter** indicates the site where RNA polymerase should begin transcribing. RNA polymerase synthesizes a single-stranded RNA using the same complementary base-pairing rules as for DNA, except that the nucleotide T is replaced by U; this process continues until a **terminator** sequence is reached. Within the resulting mRNA **transcript** is the **coding sequence** of the gene (Figure 1.8), beginning with a **start codon** (AUG) and ending with a **stop codon** (UGA, UAA, or UAG). The ribosome uses these sequences as its start and stop signals for translation; each three-nucleotide **codon** between them represents an amino acid. The protein is then a chain of amino acids (which later folds into a three-dimensional structure), each of which can be represented by a three-letter or one-letter code; thus, like the nucleic acids, proteins can be simplified for bioinformatic purposes to a string of letters. DNA and protein sequences retrieved from genomic databases use these strings as shorthand representations of complex biological molecules.

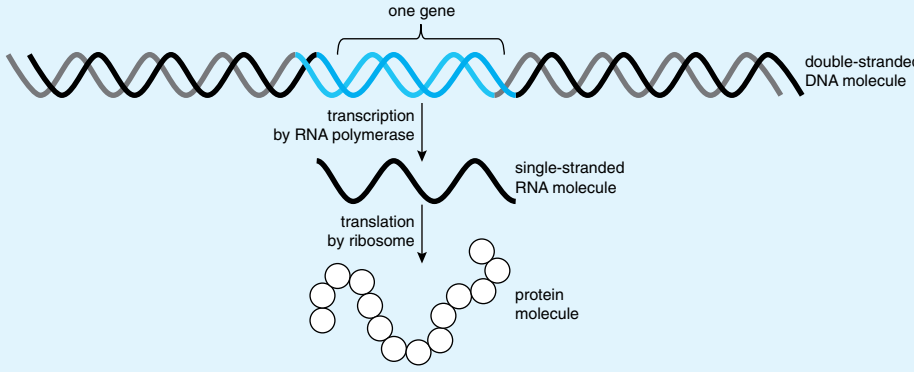

Figure 1.6 The process of gene expression, or the "central dogma of molecular biology." Information for one gene is copied from one strand of DNA to produce mRNA (transcription); this information is then decoded to produce the corresponding protein (translation).

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

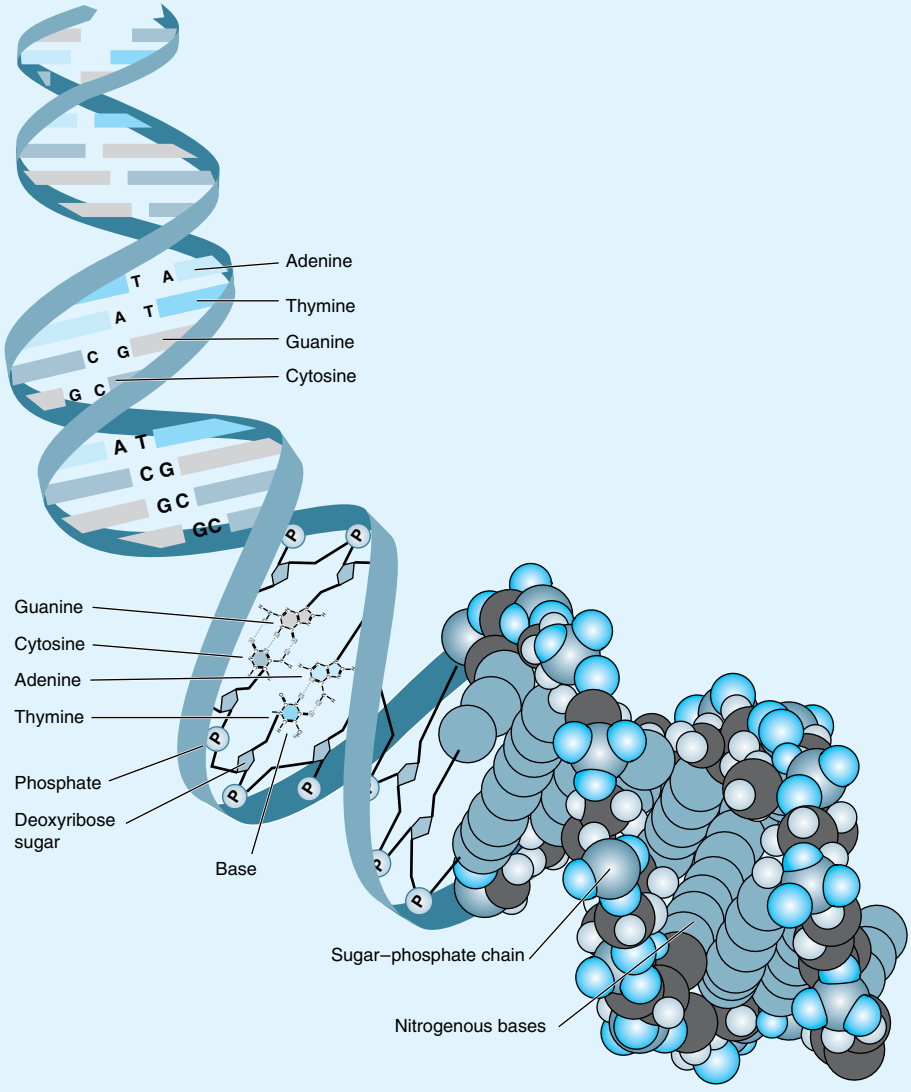

Figure 1.7 The two strands of base-paired nucleotides that make up a DNA molecule.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

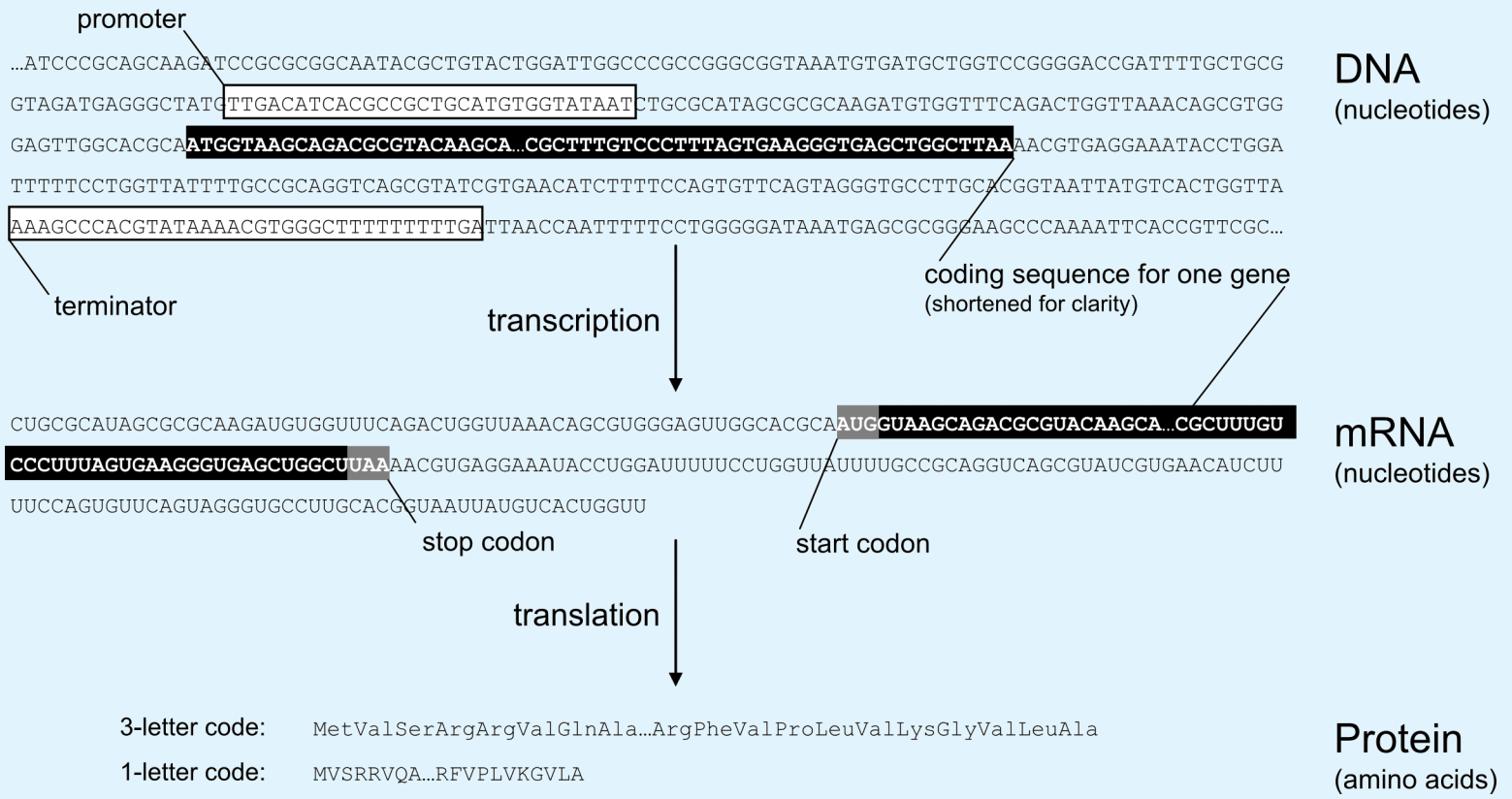

Figure 1.8 DNA, RNA, and protein can be represented by strings that are easy to process computationally. Here, the same process of gene expression as in Figure 1.3 is shown using letters to represent the nucleotides that make up DNA and RNA and the amino acids that make up proteins.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.

References and Supplemental Reading

Genome-Wide Association Study Identifying New Genome Regions Associated with Parkinson Disease

Do, C. B., J. Y. Tung, E. Dorfman, A. K. Kiefer, E. M. Drabant, U. Francke, J. L. Mountain, S. M. Goldman, C. M. Tanner, J. W. Langston, A. Wojcicki, and N. Eriksson. 2011. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. **7**:e1002141.

Overview of the UCSC Genome Browser

Zweig, A. S., D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. 2008. UCSC genome browser tutorial. Genomics **92**:75–84.

Overview of GenBank

Benson, D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. GenBank. Nucleic Acids Res. **40**:D48–D53.

NHGRI Catalog of Published GWAS

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. **106**:9362–9367.

To Learn More About GWAS

Pearson, T. A., and T. A. Manolio. 2008. How to interpret a genome-wide association study. J. Am. Med. Assn. **299**:1335–1344.

© Jones & Bartlett Learning, LLC. NOT FOR SALE OR DISTRIBUTION.