Brutlag, D. L. (1994). Understanding the Human Genome. In Leder, P., Clayton, D. A. and Rubenstein, E. (Ed.), *Scientific American: Introduction to Molecular Medicine* (pp. 153-168). New York NY: Scientific American Inc.

Understanding the Human Genome

Douglas L. Brutlag, Ph.D.

There are many compelling reasons to determine the complete genetic information of the human organism. The genome contains the history of human evolution and specifies the mechanism of human development; all humanity's physical capabilities and deficiencies are encoded in the genome. It is no wonder that Walter Gilbert, Nobel laureate and developer of one of the first methods for determining DNA sequence, has said that "sequencing the human genome is like pursuing the holy grail." Less poetically but perhaps more prophetically, Lee Hood has observed, "The sequence of the human genome would be perhaps the most useful tool ever developed to explore the mysteries of human development and disease."

For these promises to materialize, it will be necessary not only to determine the entire sequence of the human genome [see Chapter 6] but also to understand it. There is a common misconception that because the genetic code is known, the genetic information in the genome will be immediately understood. The genetic code describes only the relation between the sequences of bases in DNA that encode proteins and the sequence of the amino acids in those proteins [see Table 1]. Because less than five percent of the human genome encodes proteins, the genetic code will not help biologists to understand the function of the remaining 95 percent of the genome.

Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	
GCA GCC GCG GCT	CGA CGC CGG CGT AGA AGG	GAC GAT	AAC AAT	TGC TGT	GAA GAG	CAA CAG	GGA GGC GGG GGT	CAC CAT	ATA ATC ATT	
Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**	**Stop**
CTA CTC CTG CTT TTA TTG	AAA AAG	ATG	TTC TTT	CCA CCC CCG CCT	TCA TCC TCG TCT AGC AGT	ACA ACC ACG ACT	TGG	TAC TAT	GTA GTC GTG GTT	TAG TAA TGA

Table 1 The Genetic Code: DNA Codons Corresponding to Individual Amino Acids
The Genetic Code. This table shows the mapping of DNA triplets to amino acids present in protein coding regions of genomes. The DNA is, in fact, first transcribed into RNA and then translated into protein. In addition to the twenty amino acids, there are three codons that specify translation stop signals.

A good example of this dilemma is the recently cloned gene for cystic fibrosis.[1-3] The region of the gene transcribed into RNA is more than 250,000 base pairs long. This large RNA is processed by splicing to yield a contiguous protein-coding region in the mature messenger RNA that is only 6,500 bases long (2.6 percent of the gene). More than 97 percent of the genetic information in the cystic fibrosis gene is discarded during this splicing process, and then only 4,440 bases in the messenger RNA are actually translated into protein. The protein itself is rather large, measuring 1,480 amino acids in length.[4] Another, less extreme example of the difference between gene length and the coding regions is the human α-globin gene and its RNA and protein products [see Figure 1].

Human α-Globin Gene and its Products

```
GCCGCGCCCCGGGCTCCGCGCCAGCCAATG AGCGCCGCCCGGCCGGGCGTGCCCCCGCGC CCCAAGCATAAACCCTGGCGCGCTCGCGGC CCGGCACTCTTCTGGTCCCCACAGACTCAG    120
                                                                                                 acucuucuggucccacagacucag

AGAGAACCCACCATGGTGCTGTCTCCTGCC GACAAGACCAACGTCAAGGCCGCCTGGGGT AAGGTCGGCGCGCACGCTGGCGAGTATGGT GCGGAGGCCCTGGAGAGGTGAGGCTCCCTC    240
agagaacccaccauggugcugucuccugcc gacaagaccaacgucaaggccgccuggggu aaggucggcgcgcacgcuggcgaguauggu gcggaggcccuggagaggugaggcucccuc
            MetValLeuSerProAla AspLysThrAsnValLysAlaAlaTrpGly LysValGlyAlaHisAlaGlyGluTyrGly AlaGluAlaLeuGluArg

CCCTGCTCCGACCCGGGCTCCTCGCCCGCC CGGACCCACAGGCCACCCTCAACCGTCCTG GCCCCGGACCCAAACCCCACCCCTCACTCT GCTTCTCCCCGCAGGATGTTCCTGTCCTTC    360
cccugcuccgacccgggcuccucgcccgcc cggacccacaggccacccucaaccguccug gccccggacccaaaccccaccccucacucu gcuucucccccgcaggauguuccuguccuuc
                                                                                                 MetPheLeuSerPhe

CCCACCACCAAGACCTACTTCCCGCACTTC GACCTGAGCCACGGCTCTGCCCAAGTTAAG GGCCACGGCAAGAAGGTGGCCGACGCGCTG ACCAACGCCGTGGCGCACGTGGACGACATG    480
cccaccaccaagaccuacuucccgcacuuc gaccugagccacggcucugcccaaguuaag ggccacggcaagaagguggccgacgcgcug accaacgccguggcgcacguggacgacaug
ProThrThrLysThrTyrPheProHisPhe AspLeuSerHisGlySerAlaGlnValLys GlyHisGlyLysLysValAlaAspAlaLeu ThrAsnAlaValAlaHisValAspAspMet

CCCAACGCGCTGTCCGCCCTGAGCGACCTG CACGCGCACAAGCTTCGGGTGGACCCGGTC AACTTCAAGGTGAGCGGCGGGCCGGGAGCG ATCTGGGTCGAGGGGCGAGATGGCGCCTTC    600
cccaacgcgcuguccgcccugagcgaccug cacgcgcacaagcuucggguggacccgguc aacuucaaggugagcggcgggccgggagcg aucugggucgaggggcgagauggcgccuuc
ProAsnAlaLeuSerAlaLeuSerAspLeu HisAlaHisLysLeuArgValAspProVal AsnPheLys

CTCTCAGGGCAGAGGATCACGCGGGGTTGC GGGAGGTGTAGCGCAGGCGGCGGCGCGGCT TGGGCCGCACTGACCCTCTTCTCTGCACAG CTCCTAAGCCACTGCCTGCTGGTGACCCTG    720
cucucagggcagaggaucacgcgggguugc gggagguguagcgcaggcggcggcgcggcu ugggccgcacugacccucuucucugcacag cuccuaagccacugccugcuggugacccug
                                                                 LeuLeuSerHisCysLeuLeuValThrLeu

GCCGCCCACCTCCCCGCCGAGTTCACCCCT GCGGTGCACGCTTCCCTGGACAAGTTCCTG GCTTCTGTGAGCACCGTGCTGACCTCCAAA TACCGTTAAGCTGGAGCCTCGGTAGCCGTT    840
gccgcccaccucccgccgaguucaccccu gcggugcacgcuucccuggacaaguuccug gcuucugugagcaccgugcugaccuccaaa uaccguuaagcuggagccucgguagccguu
AlaAlaHisLeuProAlaGluPheThrPro AlaValHisAlaSerLeuAspLysPheLeu AlaSerValSerThrValLeuThrSerLys TyrArg

CCTCCTGCCCGCTGGGCCTCCCAACGGGCC CTCCTCCCCTCCTTGCACCGGCCCTTCCTG GTCTTTGAATAAAGTCTGAGTGGGCGGCAG CCTGTGTGTGCCTGGGTTCTCTCTGTCCCG    960
ccuccugcccgcugggccucccaacgggcc cuccucccccuccuugcaccggcccuuccug gucuuugaauaaagucugagugggcggc

GAATGTGCCAACAATGGAGGTGTTTACCTG TCTCAGACCAAGGACCTCTCTGCAGCTGCA TGGGGCTGGGGAGGGAGAACTGCAGGGAGT ATGGGAGGGGAAGCTGAGGTGGGCCTGCTC   1080

AAGAGAAGGTGCTGAACCATCCCCTGTCCT GAGAGGTGCCAGCCTGCAGGCAGTGGC                                    1137
```

Figure 1 The human α-globin gene is 1137 base pairs in length (top line). The primary transcript begins at base 96 and continues to 928 (833 bases) and is shown in lower case below the gene (middle line). The processed messenger RNA is missing two introns, including bases 229 to 345 and 550 to 690 (shaded), and it is only 575 bases long. The α−globin protein is 141 amino acids in length and is shown below its coding region (bottom line, blue).

Even for the five percent of the DNA that encodes protein, one cannot fully predict the structure and function of the protein product merely from its sequence. The goal of this chapter is to show what kinds of information can be derived from genetic sequences and to describe the methods currently available for Interpreting the genetic message. Most current methods involve comparing new sequences with preexisting ones, discovering structure and function by homology rather than through a true understanding of the biologic principles underlying structure and function. Before reviewing these methods, it will be useful to consider just why an understanding of gene sequences is critical to medicine.

Medical Benefits from Sequencing the Genome

Many of the immediate medical benefits to be derived from sequencing the human genome do not involve any real understanding of how genes work or of the proteins they encode. Among these benefits is the development of diagnostic DNA probes and therapeutic products for inherited disease. Knowing the difference between the DNA sequence of a normal gene and the gene responsible for a disease, one can design a DNA probe that can detect the difference and thus diagnose the disease. The classic example of this is the gene for the most common form of cystic fibrosis, which differs from its normal counterpart in that it lacks three DNA bases [see Figure 2]. In principle, DNA probes can form the basis of a clinical diagnosis for any inherited disease. If the sequence of an infectious agent is known, it is possible to develop a sensitive clinical test for that infectious disease as well.

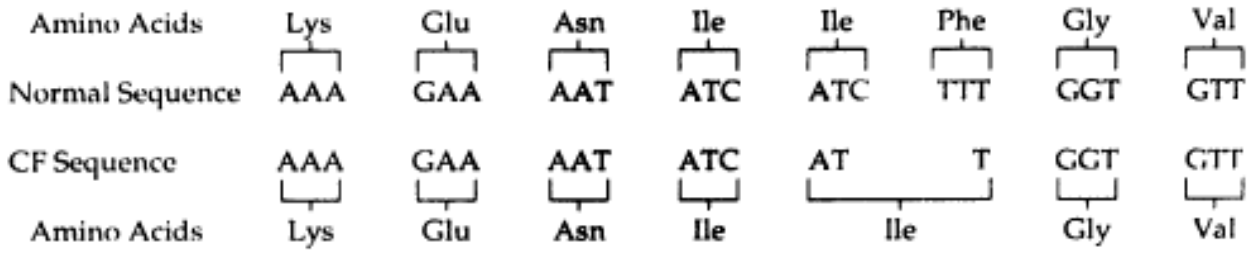

Figure 2 Shown is the 3 bp deletion in cystic fibrosis. This mutation consists of the deletion of CTT in the tenth exon of the gene. Loss of 3 bp maintains the reading frame, so that only a single phenylalanine is lost from the protein. The frequency of this mutation in cystic fibrosis patients ranges from as low as 30 percent in parts of southern Europe to 95 percent in Denmark. In Northern Europe it is 75 percent, and in the US population it is about 70 percent.

DNA probes have advantages over such diagnostic tests as radioimmune assays or bacterial cultures. The primary advantage is that the presence of a specific DNA sequence underlying an inherited or infectious disease is the most fundamental indicator of the disease; all other manifestations are secondary. This is the major reason that most biomedical problems are being attacked primarily at the level of the molecular biology. Nowadays, scientists who want to understand, detect, or treat a disease generally attempt first to isolate the gene or genes responsible for that disease: isolation of the defective gene or genes gets at the root of the problem.

Two other major advantages of detecting disease with DNA probes are that the methods are general and that they can readily be automated. Once a probe for a disease is developed, that disease can be detected by the same clinical method as any other disease. This means that a single technology can potentially be applied in the clinic to detect any of the more than 5,000 known inherited diseases—once there is a probe for each of them) Because the methods are general and easily automated, they are ideal for routine analysis of large numbers of samples.

The generality of DNA diagnostics, coupled with the ability to automate the diagnostic tests, means that clinical screens for thousands of diseases can be performed simultaneously. A It is the immediate medical value of being able to identify DNA probes for all known medical diseases that primarily motivates the Human Genome Project.

To be sure, the ability to diagnose disease (especially inherited disease) without the ability to cure or treat it leads to numerous social and ethical problems.[7] For example, insurance companies have been free to increase rates for those in high-risk groups. When insurance companies define high-risk groups on the basis of Inheritance, they are in effect holding individuals responsible for their genetic makeup. The social stigma associated with certain genetic differences can also, like more visible phenotypes, lead to discrimination. The most frustrating effects of genetic diagnosis would occur with diseases that have a clearly debilitating or fatal effect but for which there Is no hope of a cure or therapy. That is why it is important not only to know the sequence of the human genome but also to understand it well enough to devise cures and rational, inexpensive therapies.

Another early result of having the complete sequence of the human genome will be the general availability of many natural products produced in the human body. Many inherited diseases involve a deficiency of one or more protein products. Having the genetic information in hand for whatever is produced in the body should make it possible to supply any missing products as therapeutic agents. Already many natural therapeutic agents—hormones, antibodies, and so on—can be replaced by products of biotechnology [see Chapter 11]. With the complete sequence of the human genome known, the development of these natural products into therapeutic agents will be greatly simplified and accelerated. Given proper licensing policies, it should be possible even to keep the cost of such treatment within the reach of those in need. Another longer-term benefit will be the use of the DNA itself as the therapeutic agent in gene therapy [see Chapter 12].

Understanding Genetic Information

Before rational therapies can be developed, one needs to understand the action of the altered gene. A major benefit of determining the complete sequence of the human genome will be in the knowledge it provides about the evolution, development, and functioning of the human organism. Again, the classic example Is the level of understanding achieved for cystic fibrosis. Before the isolation of the gene responsible for this most common inherited disease, it was suspected that cystic fibrosis affected the ion balance in secretory cells. That being the case, cystic fibrosis could equally well have resulted from a defect In any gene that affects or controls the expression of any ion-channel gene. The isolation of the gene established that the defect is in the gene encoding an ion channel that regulates the secretion of chloride ions from the apical membrane of secretory cells in a highly regulated fashion. The sequence of the gene revealed that it encodes a membrane protein that apparently can bind the regulatory compound cyclic AMP, which controls the passage of chloride through the membrane.

By introducing the gene for a normal chloride channel into cells from cystic fibrosis patients, scientists have been able to restore normal chloride and water balance to these cells in vitro. This not only proves that the defective gene Is the major cause of the disease hut also gives hope that it will be possible to introduce the normal gene into the tissues of afflicted individuals to correct the defect. The defective gene for cystic fibrosis has been introduced into laboratory mice, creating an effective animal model of the disease for the

first time.”’0 Clinical trials In which the normal gene is being Introduced into patients with cystic fibrosis are under way.

The Flow of Genetic Information

How do molecular biologists determine that a given gene can encode a protein and then how can they deduce the function of that gene or protein from its sequence? Even before it was possible to clone and sequence individual genes, the primary goal of molecular biology was to understand the flow of genetic information from DNA to the phenotype:

DNA—>RNA—>Protein-—>Function

Biologists have studied the maintenance of genetic information, its replication, and its transformation into different forms (RNA, proteins, and phenotype) for more than 40 years. Biophysicists have studied the structures of the various molecules, biochemists have isolated specific proteins and enzymes and studied their functions, and molecular biologists have dissected the mechanisms and regulation of each step in the flow of information.

Now that there is direct access to the genetic information itself, one can investigate just how the genetic information manifests itself during its series of transformations from genome to phenotype. One examines the flow of information from the viewpoint of an information scientist:

Genetic	—>	Molecular--—>	Biochemical—>	Biological
Information		Structure	Function	Behavior

This pathway represents a shift from the classic paradigm of information flow in molecular biology. New procedures are being developed for predicting molecular structure beginning with genetic information. The first step outlined above, converting sequence information into three-dimensional structural information, is a very active field of scientific investigation [see Chapter 3]. There are many methods for performing this prediction. One method involves predicting structure from physical principles. If it is assumed that all the forces between atoms in a molecule, including bond energies and electrostatic attraction and repulsion, are known, then it is possible to calculate the three-dimensional arrangement of atoms that has the lowest energy. This method of predicting molecular structures by minimizing their overall energy is termed molecular dynamics and requires the use of very powerful supercomputers. Other methods of predicting molecular structure involve determining what kinds of amino acid sequences or patterns of amino acids are found in each of the known protein structures. Certain amino acids, such as leucine and alanine, are very common in α-helical regions of proteins, whereas other amino acids, such as proline, are rarely if ever found in α helices. Using patterns of amino acids or rules based on these patterns, one can attempt to predict where helical regions will occur in proteins whose structure is unknown. This method is an example of a larger field of automated learning. The approach involves examining the sequences of

many proteins of known structure in an effort to infer rules or patterns that can be applied to novel protein sequences to predict their structure.

Predicting the function of a molecule from its structure has long been the domain of biophysicists, whereas predicting phenotype from biochemical functions has been the domain of biochemists and geneticists. These two steps in the flow of genetic information (i.e., predicting biochemical function and phenotype) are attacked by numerous methods from information science, including simulation.

The transition in molecular biology to the new paradigm for studying the flow of genetic information parallels biology's switch from being an observational science, limited primarily by the ability to make observations, to being a data-bound science limited by its practitioners' ability to understand large amounts of information derived from observations. This change is a natural one in the maturation of scientific fields, which generally progress from observation to theory and simulation as data—and understanding of the data—increase.

Most problems in medicine and biology are now being attacked with molecular methods. More than 60 percent of all National Institutes of Health (NIH) grants depend on molecular methods of cloning, mapping, and sequencing regardless of the problem being addressed. More than 65 percent of all articles indexed by the National Library of Medicine contain molecular biology subject headings. Whether scientists are studying cancer or allergies, aging or infectious disease, they usually apply molecular methods.

The reason for this switch to a molecular approach is obvious from an examination of either pathway for the flow of genetic information outlined above. The fundamental cause of any inherited disease is written in the genome. Even an agent of infectious disease is often best detected by a DNA diagnostic probe, because of its greater sensitivity. Studies at any other level are always secondary and subject to interpretation, whereas DNA diagnostics based on an isolated gene defect are unambiguous. The fundamental nature of the genetic information and the ease with which this information can be obtained make the molecular approach the preferred one in most cases. The benefit of this approach is the accumulation of large amounts of genetic information; the problem is that only a small fraction of this information is understood.

Another way of emphasizing the imperative to interpret the genetic information is to compare the various ways of unraveling the classical pathway of the flow of information. In the past 100 years, biochemists have isolated and characterized about 10,000 enzymes to the point that one can understand their effect on physiology. In the past 50 years, biophysicists have determined the structure of only 1,000 molecules with the degree of accuracy needed to predict their function. Yet in only the past 20 years, scientists have cloned and completely sequenced more than 140,000 genes, of which 20 percent are human. Had all the laboratories of the world been working exclusively on human genes during this period, they would have sequenced the expected number of human genes. This rate of accumulation of genetic information will increase exponentially with the added expenditure of time and money occasioned by the human genome project. To obtain the most benefit from this investment, it will be necessary to predict the structure, function, and behavior resulting from a particular genetic sequence. Hence it is essential to learn

from the few examples that are well understood, so that valuable laboratory resources can he focused on novel sequences, structures, and functions.

Challenges in Understanding Genetic Information

There are many challenges in trying to interpret genetic information. The first is that this information is highly redundant: many sequences encode the same function or message. There are more than 650 globin sequences in the protein-sequence databases, all with very similar structure. (It is much like this in the case of languages. The English language is notorious for its numerous synonyms, each with its associated nuances and hidden implications.) Of course, the genetic code itself is highly redundant. On average, three different codons (DNA sequences three bases long) encode each amino acid [See Table 1]. (Each triplet in the genetic code, of which there are 64, can represent one of 20 amino acids or the termination of translation; 64 codons divided by 21 meanings is approximately 3.05 codons per meaning. Two amino acids—methionine and tryptophan—are specified by only one codon, but several amino acids can be specified by as many as six codons.) Given the average of three codons per amino acid, one can predict that a protein as short as insulin (51 amino acids) might be encoded by as many as $3^{51} = 10^{24}$ different DNA sequences, each encoding precisely the same protein. Fortunately, nature does not fully exploit the available redundancy. Most insulin gene sequences have derived from other insulin sequences in evolution; most natural sequences are related to other preexisting natural sequences rather than being created anew. Molecular biologists can apply this sequence similarity to derive conclusions about structure and function on the basis of homology.

The most important challenge in understanding genetic information is one briefly alluded to above: genetic information is one-dimensional, but biological molecules are three-dimensional. The information for the three-dimensional nature of DNA, RNA, and proteins must be determined by the one-dimensional sequence of their component nucleotides or amino acids, because when these molecules are denatured (i.e., their native structure completely disrupted) in the laboratory, they can renature and regain their normal three-dimensional structure." The implication is that the tertiary structure of molecules is encoded in their primary sequence. But how? It is this structural code that biologists must next determine in order to understand the genetic message.

Unfortunately, most of the methods described below that are applied to the analysis of gene sequences do not capture the structural information directly. Instead, they look for evolutionarily related sequences and make the assumption that if two DNA segments are related in sequence, they are probably related in structure or function. Although this assumption is usually justified, the methods for detecting similarities at the sequence level are not sensitive enough to detect all structurally or functionally related sequences.

Methods for Understanding Genetic In formation

There are three primary methods for analyzing sequences for structure or function: consensus sequences, weight matrices, and sequence alignment.

CONSENSUS SEQUENCES

The first of these methods discovers short, highly conserved sequence patterns. Such sequences, usually called motifs or consensus sequences, are determined by aligning a large number of sequences of common function and looking for conserved positions—that is, for positions at which the same amino acids (or nucleotide bases) are present in most, or at least very many, of the sequences. Once the conserved positions are identified, the motif thus defined can serve as a kind of probe: one can search a database or some new sequence, testing for the presence of the motif.[12]

The concept of a motif is a very powerful one, and databases of motifs have been compiled. The best known is the Prosite database, which currently contains more than 1,000 such patterns." A very useful method for determining the function of a newly sequenced protein is to search it for each and every one of these motifs. With more than 1,000 motifs, there is often one that will match the new protein and thus give a clue as to its structure or function. When this method was applied to the protein sequence of the cystic fibrosis gene product, two motifs were identified that aided in the discovery of two sites on the protein that were involved in binding nucleotides (nucleotide binding folds, or NBFs).

One of the main problems with consensus sequences is the difficulty in striking a balance between their specificity and their precision. In order to eliminate as many false hits (sequences that match the pattern but are known not to be functionally or structurally related) as possible, one often has to make the consensus sequence very specific. Often, however, the motif is made so specific that it misses even many of the known examples of a motif. A motif can often be very specific (finding no matches that are not established examples of the structure or function) but not very precise (missing many other examples). Or it can be very precise (finding all the known examples) but not very specific, in that it also matches many unrelated sequences. One way around this dilemma is to classify a set of protein sequences into narrow subgroups and then to have multiple motifs, one for each subclass or subgroup. Even with this approach, it is often difficult to extract a highly conserved consensus sequence because of the high degree of ambiguity in biologic sequences.

Another fundamental problem with consensus sequences is their discrete nature. A test sequence either matches or does not match the consensus sequence; there is no concept of the degree or probability of matching between a test sequence and a consensus sequence. It is this weakness that has led to the development of probabilistic methods that can generate a more powerful representation of a series of aligned sequences.

WEIGHT MATRICES AND PROFILES

Because of the extreme redundancy of genetic sequences, many proteins having a common structure and function may have very few, if any, amino acids in common. A good example of this is the set of prokaryotic DNA-binding proteins containing a so-called helix-turn-helix motif, 22 amino acids in length, that allows them to bind to DNA in a sequence-specific manner.[4] The motif is called a helix-turn-helix motif because the protein assumes a characteristic structure involving a short a-helix, followed by a 90-degree turn of three amino acids, followed by another short helix.) Examination of the linear amino acid sequences reveals no consensus sequence that could reliably distinguish these proteins from other proteins [see Table 2].

Table 2 composition of 22—Amino Acid Sequences Corresponding to Helix-Turn-Helix Motif in 19 Prokaryotic Proteins

Sequence	Helix										Turn			Helix							
RCRO$LAMBD	F	G	Q	T	K	T	**A**	K	D	L	**G**	V	Y	Q	S	A	**I**	N	K	A	I	H
RCRO$BP434	M	T	Q	T	E	L	**A**	T	K	A	**G**	V	K	Q	Q	S	**I**	Q	L	I	E	A
RCRO$BPP22	G	T	Q	R	A	V	**A**	K	A	L	**G**	I	S	D	A	A	**V**	S	Q	W	K	E
RPC1$LAMBD	L	S	Q	E	S	V	**A**	D	K	M	**G**	M	G	Q	S	G	**V**	G	A	L	F	N
RPC1$BP434	L	N	Q	A	E	L	**A**	Q	K	V	**G**	T	T	Q	Q	S	**I**	E	Q	L	E	N
RPC1$BPP22	I	R	Q	A	A	L	**G**	K	M	V	**G**	V	S	N	V	A	**I**	S	Q	W	E	R
RPC2$LAMBD	L	G	T	E	K	T	**A**	E	A	V	**G**	V	D	K	S	Q	**I**	S	R	W	K	R
LACR$ECOLI	V	T	L	Y	D	V	**A**	E	Y	A	**G**	V	S	Y	Q	T	**V**	S	R	V	V	N
CRP$ECOLI	I	T	Q	Q	E	I	**G**	Q	I	V	**G**	C	S	R	E	T	**V**	G	R	I	L	K
TRPR$ECOLI	M	S	Q	R	E	L	**K**	N	E	L	**G**	A	G	I	A	T	**I**	T	R	G	S	N
RPC1$CPP22	R	G	Q	R	K	V	**A**	D	A	L	**G**	I	N	E	S	Q	**I**	S	R	W	K	G
GALR$ECOLI	A	T	I	K	D	V	**A**	R	L	A	**G**	V	S	V	A	T	**V**	S	R	V	I	N
Y77$BPT7	L	S	H	R	S	L	**G**	E	L	Y	**G**	V	S	Q	S	T	**I**	T	R	I	L	Q
TER3$ECOLI	L	T	T	R	K	L	**A**	Q	K	L	**G**	V	E	Q	P	T	**L**	Y	W	H	V	K
VIVB$BPT7	D	Y	Q	A	I	F	**A**	Q	Q	L	**G**	G	T	Q	S	A	**A**	S	Q	I	D	E
DEOR$ECOLI	L	H	L	K	D	A	**A**	A	L	L	**G**	V	S	E	M	T	**I**	R	R	D	L	N
RP32$BACSU	R	T	L	E	E	V	**G**	K	V	F	**G**	V	T	R	E	R	**I**	R	Q	I	E	A
Y28$BPT7	E	S	N	V	S	L	**A**	R	T	Y	**G**	V	S	Q	Q	T	**I**	C	D	I	R	K
IMMRE$BPPH	S	T	L	E	A	V	**A**	G	A	L	**G**	I	Q	V	S	A	**I**	V	G	E	E	T

The motif is, however, recognizable when a quantitative approach is taken. The frequency with which each amino acid appears at each position is determined [see Table 3]. These numbers are then used to calculate the likelihood of finding each amino acid at each position. The frequency matrix is converted to a traditional weight matrix by converting every number in the matrix to a measure of the probability of occurrence of each acid (rather than its frequency). These so-called weight matrices can be applied to measure the likelihood that any given sequence 22 amino acids long is related to the helix-turn-helix family by merely multiplying the likelihoods of each amino acid in a test sequence. A modification of this method, referred to as profiles, allows one to estimate the probability that any amino acid will appear in a specific position—even if a particular amino acid has never been observed at some positions in the set of known examples.[8]

	Position																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	2	1	3	13	10	12	67	4	13	9	1	2	4	3	6	15	4	4	4	11	0	10
R	7	5	8	9	4	0	1	16	7	0	1	0	1	16	6	6	0	11	28	3	0	16
N	0	8	0	1	0	0	0	2	1	1	10	0	7	1	3	1	0	4	8	0	1	11
D	0	1	0	1	13	0	0	12	1	0	4	0	1	2	0	0	0	0	1	1	0	3
C	0	0	1	0	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0	1	0	0
Q	1	1	21	8	10	0	0	7	6	0	0	2	1	17	7	7	0	2	12	5	2	4
E	2	0	0	9	21	0	0	15	7	3	3	0	1	6	11	0	0	2	0	1	13	6
G	9	7	1	4	0	0	8	0	0	0	46	0	6	0	7	1	0	3	1	1	0	4
H	4	3	1	1	2	0	0	2	2	0	5	0	3	3	0	2	0	2	4	5	0	2
I	10	0	11	1	2	10	0	4	9	3	0	16	0	2	0	1	26	1	0	8	16	0
L	16	1	17	0	1	31	0	3	11	24	0	14	0	2	0	1	21	1	1	12	20	0
K	3	4	5	10	11	1	1	13	10	0	5	2	1	4	1	1	0	1	8	4	5	14
M	7	1	1	0	0	0	0	0	5	7	1	8	0	0	2	0	2	0	0	2	0	1
F	4	0	3	0	0	4	0	0	0	10	0	0	0	0	1	0	0	1	1	1	11	0
P	0	6	0	1	0	0	0	0	0	0	0	0	1	12	7	0	0	0	0	0	0	3
S	1	17	0	8	3	1	3	0	2	2	2	0	37	1	24	5	0	29	3	0	1	3
T	5	22	3	11	1	5	0	2	2	2	0	5	16	4	2	38	0	4	1	0	4	3
W	2	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2	10	0	0
Y	1	0	4	2	0	1	0	0	2	4	0	1	1	2	0	2	0	15	5	7	0	0
V	6	3	1	1	2	15	0	0	2	12	0	28	0	5	3	0	27	0	1	8	7	0

Table 3 Frequency of Occurrence of Each Amino Acid at Each Position in 80 Examples of Helix-Turn-Helix Motif.

These weight matrices or profiles are effective for locating signals in DNA as well as structural motifs in proteins. Indeed, the first application of these probabilistic methods was the detection of promoters and splice sites in DNA by means of weight matrices.'9

SEQUENCE ALIGNMENT

The most common way to examine a new gene or protein for its biologic function is simply to compare its sequence with all known DNA or protein sequences in the public databases and note any strong similarities.[20-22] The particular gene or protein that has just been determined will of course not be found in the databases, but a homologue from another organism or a gene or protein having a related function may be found. In either case the evolutionary similarity implies a common ancestor and hence a common function. This method becomes more and more successful as the databases grow larger and as the sensitivity of the search procedure increases.

It is for these reasons that it is critical to have available both the most sensitive search procedure and access to the most up-to-date databases possible. Since 1983, the National Institutes of Health has supported the Genbank DNA Database,[23,24] and the European Molecular Biology Laboratory the EMBL DNA Library [25,26]; both databases collect and disseminate all published DNA sequences to the scientific community. These computer resources, together with the younger DNA Database of Japan, accept sequences electronically from the molecular biology community, verify them, and then redistribute them daily to researchers around the world, again via electronic networks. A new

sequence is generally on file in these databases within a day or two of its publication. Similar databases of protein sequences are also maintained at various sites.[27,28]

In addition to up-to-date databases, It Is critical to be able to compare a query sequence against all the database sequences with methods that allow a flexibility of matching commensurate with the redundancy of the genetic code.[29,30] A simple method that merely lines up identical amino acids, for instance, will not detect similarities between similar genes or proteins in organisms as distantly related as plants and animals. (Plant and animal hemoglobins, for example, have less than 10 percent identity of amino acids in such alignments.) By assigning a measure of similarity to each pair of amino acids, however, and then adding up these pairwise scores for the entire alignment, it is possible to detect highly significant similarities between even distantly related proteins. Although related proteins may not have identical amino acids aligned, they usually do have chemically similar or replaceable amino acids in similar positions. In the type of scoring that is usually applied to such alignments,[31] amino acid pairs that are identical or chemically similar are given positive scores, and pairs of amino acids that are not related are assigned negative similarity scores. Negative penalties usually have to be designated for the introduction of insertion / deletion gaps. Use of such matrices markedly improves the sensitivity of a database search.

Many computer programs have been developed for finding the most similar sequences in a database by applying the scoring procedures described above. Because making tens of thousands of alignments of a new sequence with every known sequence is computer-intensive, many methods are mere approximations to the complete alignment method.[32]

Other approaches rely on massively parallel computers or supercomputers to carry out database search and alignment programs. The amino acid sequence of the cystic fibrosis transmembrane conductance regulator protein (CFFR) has been compared [see Figure 3] with the sequences of 26,706 proteins in a current protein database by means of a massively parallel computer. An examination of the list of the top 27 similar proteins strongly suggests that the CFI'R protein Is a membrane protein involved in secretion. There are also highly significant homologies to ATP-binding transport proteins. (Incidentally, some of the most highly significant homologies with this human protein are found for proteins from organisms as remote as Escherichia coli and yeast.) It was results such as these that led to an understanding of the function of the CFTR protein.

```
      ==========                  B L A Z E   (tm)
                ========
              ========       A  High-Performance High-Sensitivity Biological
            ========       Sequence    Similarity    Searching     Program
          ========       Utilizing a Massively  Parallel  Implementation
        ========       of  the   Dynamic  Programming  Algorithm    of
       ========       Smith  and  Waterman
      ==========  Release 1.0 - July 1992  ======================
Copyright (c) 1992 by IntelliGenetics, Inc. and MasPar Computer Corporation

SEARCH STATISTICS
Number of sequences searched:                          26,706
Number of residues in database:                     9,011,391
Query sequence length:                                  1,480
Score of query vs. itself:                              2,279
Mean score:                                                33
Standard Deviation:                                     25.47

  Sequence Name   Description                          Length Score %Match   Exp
-------------------------------------------------------------------------------
   1. CFTR_HUMAN  CYSTIC FIBROSIS TRANSMEMBRANE COND   1480  2279   100   0.000
   2. CFTR_XENLA  CYSTIC FIBROSIS TRANSMEMBRANE COND   1485  1898    83   0.000
   3. CFTR_MOUSE  CYSTIC FIBROSIS TRANSMEMBRANE COND   1476  1874    82   0.000
   4. CFTR_SQUAC  CYSTIC FIBROSIS TRANSMEMBRANE COND   1492  1798    79   0.000
   5. MDR_LEITA   MULTIDRUG RESISTANCE PROTEIN (P-GL   1548   536    24   0.000
   6. HETA_ANASP  HETEROCYST DIFFERENTIATION PROTEIN    607   247    11   0.000
   7. MDR2_MOUSE  MULTIDRUG RESISTANCE PROTEIN 2 (P-   1276   240    11   0.000
   8. STE6_YEAST  MATING FACTOR A SECRETION PROTEIN    1290   231    10   0.000
   9. MSBA_ECOLI  PROBABLE ATP-BINDING TRANSPORT PRO    582   230    10   0.000
  10. MDR3_CRIGR  MULTIDRUG RESISTANCE PROTEIN 3 (P-   1281   226    10   0.001
  11. CYAB_BORPE  CYAB PROTEIN.                         712   224    10   0.001
  12. MDR1_HUMAN  MULTIDRUG RESISTANCE PROTEIN 1 (P-   1280   218    10   0.001
  13. MDR1_CRIGR  MULTIDRUG RESISTANCE PROTEIN 1 (P-   1276   216     9   0.002
  14. MDR2_CRIGR  MULTIDRUG RESISTANCE PROTEIN 2 (P-   1276   216     9   0.002
  15. MDR3_HUMAN  MULTIDRUG RESISTANCE PROTEIN 3 (P-   1279   216     9   0.002
  16. MDR1_MOUSE  MULTIDRUG RESISTANCE PROTEIN 1 (P-   1276   214     9   0.002
  17. MDR3_MOUSE  MULTIDRUG RESISTANCE PROTEIN 3 (P-   1104   210     9   0.003
  18. LKTB_ACTAC  LEUKOTOXIN SECRETION PROTEIN.         707   203     9   0.006
  19. HLYB_ECOLI  HAEMOLYSIN SECRETION PROTEIN, PLAS    707   201     9   0.006
  20. HLY2_ECOLI  HAEMOLYSIN SECRETION PROTEIN, CHRO    707   201     9   0.006
  21. LKTB_PASHA  LEUKOTOXIN SECRETION PROTEIN.         708   199     9   0.008
  22. HLYB_PROVU  HAEMOLYSIN SECRETION PROTEIN.         707   194     9   0.013
  23. HLYB_ACTPL  HAEMOLYSIN SECRETION PROTEIN (CLYI    707   183     8   0.034
  24. PRTD_ERWCH  PROTEASES SECRETION PROTEIN PRTD.     575   178     8   0.055
  25. CVAB_ECOLI  COLICIN V SECRETION PROTEIN CVAB.     698   164     7   0.182
```

Figure 3. Illustrated is a database search for proteins homologous to the cystic fibrosis transmembrane receptor (CFTR) protein. The human CFTR protein was used as a query and compared to all the known protein sequences via the BLAZE program from IntelliGenetics, Inc., running on a massively parallel computer from MasPar Computer Corporation. The similarity scores for the top 25 protein sequences were printed out. Sequences with an expectation (far right) of 0.05 or less are considered statistically significant. Notice that match scores as low as eight percent are significant. This is because of the use of an amino acid matching matrix that allows similar amino acids to score highly.

Toward True Understanding of the Genome

To reap the most benefit from having in hand the human genome, it will be necessary to understand the meaning of the genetic sequences. The methods currently available for interpreting DNA and protein sequences largely utilize evolutionary homology. The consensus-sequence method looks for highly conserved amino acids or bases in specific locations. Weight matrix or profile methods perform the same task quantitatively. With these evolution-based methods, as this chapter has shown, much hypothetical information can be gained from the study of a single gene and protein molecule.

Yet these evolution-based methods do not give much insight into the flow of genetic information from genes to structure and to phenotype, a goal discussed above. What are truly needed are methods that can predict structure and function on the basis of physical and chemical principles. Such methods will have to embody knowledge about how proteins fold, how they mediate catalysis, how they interact, and how they determine phenotype. Research is under way along these lines. Methods based on molecular dynamics aim to predict the structure of DNA, RNA, and proteins from physical principles. Automated learning methods, including some that depend on neural networks, are directed at identifying these physical principles through an analysis of the large amounts of sequence information now available. Probabilistic networks and other statistical methods may also reveal principles of physical structure and function based on examples in the growing public databases.

Even without these sophisticated informatic methods, there is still much to gain from knowing the sequences of the human genome. At the very least it will be possible to design DNA diagnostic probes for many, if not all, inherited diseases. Genome sequences coupled with recombinant DNA technology will make it possible to synthesize and mass produce human proteins having therapeutic value. Eventually, enhanced ability to decode the genetic information will make it possible to understand the nature of disease states and to design more rational therapies in the short term and genetic therapies in the long term.

References

1. Kerem, B, Rommens, JM, Buchanan, JA, et al: Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073, 1989

2. Riordan, JR, Rommens, JM, Kerem B, et al: Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066, 1989

3. Rommens, JM, lannuzzi MC, Kerem BS, et al: IdentifIcation of tile cystic fibrosis gene: chromosome walking and jumping. Science 245:1059, 1989

4. Collins FS: Cystic fibrosis: molecular biology and therapeutic implications. .Science 256:774, 1992

5. McKusick VA: Current trends in mapping human genes. FASER / 5:12, 1991

6. Fodor SPA, Read JL, Pirrung MG, et al: Light-directed, spatially adressable parallel chemical synthesis. Science 251: 767, 1991

7. Murray TH: Ethical issues in human genome research. FASER 15:55, 1991

8. Clarke LL, Grubb RR, Gabriel SE, et al: Defective epithelial chloride transport in a gene-targeted mouse model of cystic fibrosis. Science 257:1125, 1992

9. Dorm JR, Dickinson P, Alton EW, et al: Cystic fibrosis in the mouse by targeted insertional mutagenesis. Nature 359:211, 1992

10. Snouwaert JN, Brigman KK, Latour AM, et al: An animal model for cystic fibrosis made by gene targeting. Science 257:1083, 1992

11. Anfinsen GB: Principles that govern the folding of protein chains. Science 181:223, 1973

12. Abarbanel RM, Wieneke PR, Mansfield E, et al: Rapid searches for complex patterns in biological molecules. NucleicAcids Res 12:263, 1984

13. Ilairoch A: PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. 19:2241, 1991

14. Brennan RG, Matthews BW: The helix-turn-helix DNA binding motif. J Biol Chem 264:1903, 1989

iS. Dodd IB, Egan JB: The prediction of helix-turn-helix DNA-binding regions in proteins: a reply to Yudkin. Protein Eng 2:174, 1988

16. Dodd IB, Egan JB: Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res 18:5019, 1990

17. Gribskov M, Homyak M, Edenfield J, et al: Profile scanning for three-dimensional structural patterns in protein sequences. Comput Appi Biosci 4:61, 1988

18. Gribskov M, McLachian AD, Eisenberg D: Profile analysis: dectection of distantly related proteins. Proc Nati Acad Sci USA 84:4355, 1987

19. Staden R: computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res 12:505, 1984

20. Lipman DJ, Pearson WR: Rapid and sensitive protein similarity searches. Science 227:1435, 1985

21. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444, 1988

22. Wilbur WJ, Lipman, DJ: Rapid simIlarity searches of nucleic acl(l and proteIn data banks. Proc Nati Acad Sd USA 80:726, 1983

23. Benton D: Recent changes in the GenBank on-line service. Nucleic Acids Res 18:1517, 1990

24. Cinkosky MJ, Pickett JW, Gilna P, et al: Electronic Data publishing and GenBank. Science 252:1273, 1991

25. Cameron GN: The EMBL data library. Nucleic Acids Res 16:1865, 1988

26. Stoehr PJ, Omond RA: New nucleotide sequence data on the EMil!. Pile Server. Nucleic Acids Res 17:6765, 1989

27. Bairoch A, Boeckmann B: The SWISS-PROT protein sequence data bank. Nucleic Acids Res 19:2247, 1991

28. Sidman KE, George DG, Barker WC, et al: The protein identification resource (PIR). Nucleic Acids Res 16:1869, 1988

29. Needleman SB, Wunsch CD: A general method applicable) to the search for sinillarities in the amino acid sequence of two proteins. J. Mol Biol 48:443, 1970

30. Smith TF, Waterman, M: IdentIfication of common molecular subsequences.J. Mol. Biol. 147:195, 1981

31. Schwartz, RM, Dayhoff MO: Matrices for detecting distant relationships. Atlas of Protein Structure 5(suppl 3):353, 1979

32. Brutlag, D. L., Dautricourt, J.-P., Maulik, S, et al: Improved sensitivity of biological sequence database searches. Comput Appi Biosci 6:237, 1990

Acknowledgments

Figures 1, 3 Sal Terillo.

Figure 2 Talar Agasyan. Adapted from Recombinant DNA, 2nd ed., byj. D. Watson, M. Gilman, J. Witkowski, et a!. Scientific American Books, New York, 1992. 1992 James D. Watson, Michael Gilman, Jan Wltkowski, Mark Zoller. Used by permission.

Table 2 Data from “The Helix-Turn-Helix Binding Motif,” by R. G. Brennan and B. W. Mathews, in Journal of Biological Chemistry 264:1903, 1989.