Human Reproduction Update 1996, Vol. 2, No. 2 pp. 172–191

Genetic engineering of human FSH (Gonal-F®)

C.M.Howles

Ares-Serono, 15 bis Chemin des Mines, 1202 Geneva, Switzerland

Telephone: 41 22 738 8000; Fax: 41 22 739 3043

TABLE OF CONTENTS

I. An introduction to protein biosynthesis

TABLE OF CONTENTS

Introduction

Molecules of all types abound in living cells and for our purposes they can be divided into two categories: small and large. Although the small molecules are extremely important, even vital to life, these will not be considered here—it is the large (or macromolecules) which we will concentrate on and these come in three classes: proteins, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

After water, proteins are the major constituent of cells and are involved in virtually every cellular process. In order to perform a specific function, a cell has to produce the particular proteins involved with that task. For instance, a muscle cell will produce large quantities of a protein called actin which is involved with enabling the muscle to contract. Other cells found in endocrine glands produce large amounts of certain proteins which the cell secretes into the blood. These proteins, called hormones, travel to other parts of the body where they cause other cells to function differently. An example is insulin, which is only produced in the pancreas, but causes cells all over the body to take up glucose. Another hormone, follicle stimulating hormone (FSH), interacts exclusively with two cell types, the Sertoli in the male or the granulosa cell in the female. There are many more functions that proteins serve, e.g. another important role is as an enzyme.

Like their functions, protein molecules vary tremendously. They come in different sizes and shapes but they share one thing in common. They are all made by linking together groups of smaller molecules, called amino acids. Twenty different amino acids or more are used to make up a protein, and exactly the right amino acids must be linked together in a set order to give that particular protein its characteristics.

Proteins: what are they?

Proteins are made up of amino acids: 20 amino acids are commonly used. The order of the amino acids is important in determining the structure/function of the protein. A protein has a three-dimensional (3D) structure that is pivotal for its function and interaction with receptor/substrate. A single alteration (mutation) in the amino acids sequence can render the protein inactive.

In order for many of the proteins to be able to carry out their functions correctly, they must have a particular 3D structure. For instance, with an enzyme, certain amino acids in the structure must be positioned so that they can interact properly with their specific substrate and thus catalyse a given reaction. If this positioning is faulty or disrupted, the enzyme may no longer work.

In order to obtain the unique 3D structure of a particular protein, the individual amino acids are able to rotate around the peptide bonds which form them, thereby allowing them to assume a particular configuration. The conformation of a protein is determined primarily by its amino acid sequence. In addition, for a molecule such as a gonadotrophin the tertiary structure is also shaped by the formation of disulphide bonds which link between certain amino acids that are remote from each other in the primary amino acid sequence. The correct tertiary structure of the molecule is also necessary for its interaction with its receptor.

DNA and genes

To synthesize a protein in a laboratory, a biochemist will follow a set of instructions which tell him which amino acids to add next. Inside a cell the ‘machinery’ which produces proteins must also follow a set of instructions. These instructions are carried by another macromolecule called DNA which, in mammalian cells, is found in the chromosomes which are located in the nucleus. The basic building block of DNA is a nucleotide, comprising a base, a sugar and a phosphate group. There can be four types of nucleotide, the difference being the type of base which they contain. DNA is made of two long strands coiled around each other to form a double helix (Figure 1). Most DNA is present in the cell nucleus. Each cell in the human body contains ~1.8 m of DNA. The amount of DNA in all cells of a single human would stretch out to the moon and back 8000 times!

One way of thinking of DNA is as a ‘library’ which stores all the information (in recipe form) required for the cell to successfully make the proteins necessary for normal function. The recipes (or genes) hold the information to make sure that the right amino acids are put together in a strict order so as to make a fully functioning protein. However, this is a very special library in which there is no

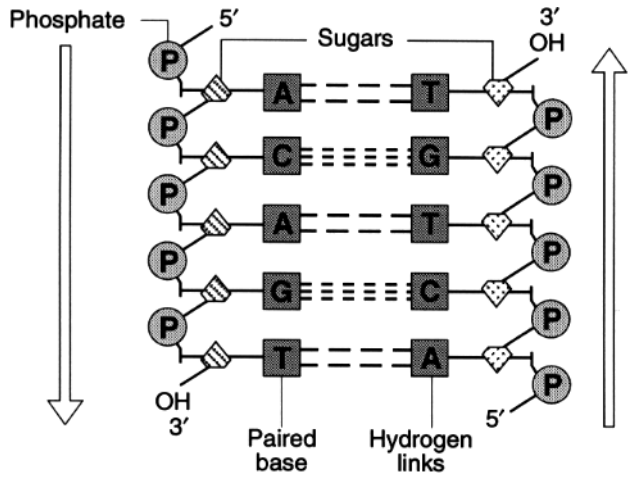

Figure 1. Representation of the two sugar backbone chains of DNA, aligned head to tail.

First Position	Second Position				Third Position
	T	C	A	G	
T	PHE	SER	TYR	CYS	T
	PHE	SER	TYR	CYS	C
	LEU	SER	stop	stop	A
	LEU	SER	stop	TRP	G
C	LEU	PRO	HIS	ARG	T
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE	THR	ASN	SER	T
	ILE	THR	ASN	SER	C
	ILE	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	T
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

Figure 2. The genetic code. For example, the codon TGT or TGC codes for the amino acid cysteine.

master index, the ‘books’ do not have a table of contents, the pages containing the information are not indicated and there are no sentences or full stops!

Even changing one amino acid (one form of mutation) can alter the nature of a protein. For example, one amino acid change to haemoglobin, a protein which carries oxygen in the blood, causes a disease called sickle cell anaemia. People with sickle cell anaemia have a gene which substitutes valine instead of glutamic acid in the sixth amino acid position of the haemoglobin molecule.

The genetic code

There are four different bases present in DNA called adenosine (A), cytosine (C), guanine (G) and thymine (T). One of these four bases is present in each nucleotide (made of a sugar, phosphate group and base). The order of amino acids in a protein is encoded in turn by the exact sequence of bases in the DNA of the gene which is responsible for that protein. Thus the sequence of bases forms the genetic information carried by the DNA.

An important aspect of this genetic code is that it always takes three nucleotides (the triplet code) in a specific order to code for a single amino acid. For example, the nucleotide sequence thymine, cytosine, guanine (TCG) will always code for the amino acid alanine. This triplet code is similar for many species, from bacteria to man (Figure 2).

Three of the triplets fail to specify any amino acid but are signals that the protein sequence is finished.

Protein biosynthesis

During protein biosynthesis, which occurs in the cell cytoplasm at specialized structures called ribosomes, the instructions to the cell are explicit: it must add the amino acid coded for by each triplet of nucleotides. In this way, not only can we produce the exact copies of the proteins we need throughout our lives but we can also pass this information on to our offspring in their genes. For this to happen the DNA molecule must be stable.

DNA carries the instructions for protein synthesis, but it does not actually participate directly in the process. Rather, an intermediate molecule called ribonucleic acid (RNA) is produced from the DNA (gene) template. RNA is a similar molecule to DNA except that it has ribose as its sugar and the thymine base is replaced by uracil. This RNA molecule (messenger RNA) is single stranded, is made up of a specific nucleotide sequence and is an exact mirror of the DNA sequence from which it was produced. Thus, the coded instructions are maintained and this process is called transcription. Transcription starts at specific sites on the DNA molecule, called promoters. As it is transcribed, the DNA helix unzips and an enzyme, RNA polymerase, travels along the exposed DNA strand, bringing in nucleotides to make mRNA from the DNA template. At a specific sequence which terminates transcription (terminator or ‘stop’ codon), the RNA polymerase detaches from the DNA strand and the mRNA is released.

Once made, mRNA travels out from the nucleus and enters a large cellular structure called a ribosome which is found in the cell cytoplasm. It is here that the nucleotides which make up the mRNA are translated into the correct sequence of amino acids required to make a protein.

The ribosomes in Figure 3 are moving along the mRNA from left to right. As they do so, small adaptor RNAs (called transfer or tRNAs) read the sequence on the mRNA and add an appropriate amino acid to the next position on

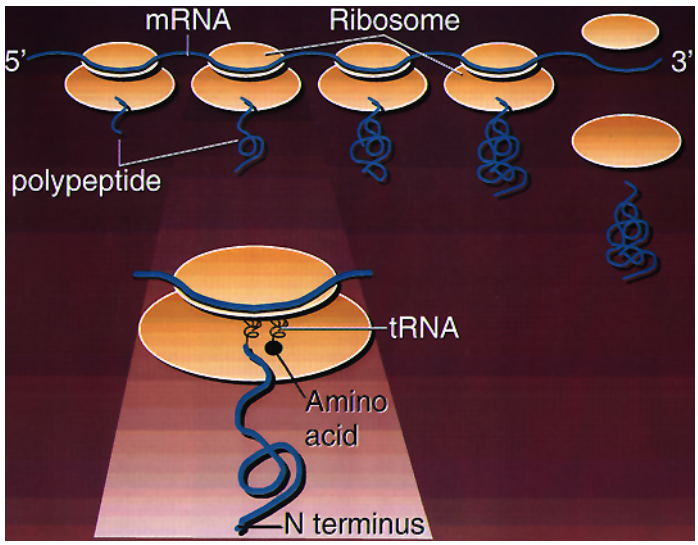

Figure 3. Protein synthesis: the role of the ribosome.

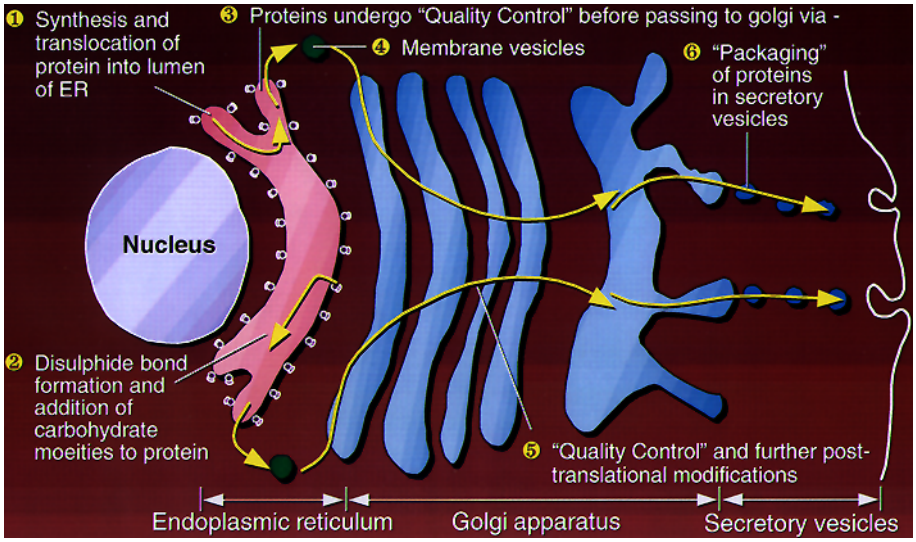

Figure 4. Synthesis and migration of secretory proteins through the cytoplasmic organelles of a eukaryotic cell.

the growing peptide chain. There are many tRNAs. Each has a loop at one end that bears a sequence of nucleotides that is complementary to a triplet on the mRNA. On its other end, the tRNA bears an amino acid. The particular amino acid that the tRNA carries corresponds to the sequence of the complementary loop near its other end. Notice that the protein is synthesized beginning at its amino acid terminal end (called the N terminus).

The ribosomes are tightly bound to the 'rough' endoplasmic reticulum (ER). Following synthesis, secretory proteins are translocated into the lumen of the rough ER, where the signal sequence is cleaved (Figure 4). Disulphide bond formation, and the addition of carbohydrates to specific asparagine acid (Asn) residues (glycosylation), also occurs within the ER. 'Quality control' of incorrectly folded or aggregated proteins is carried out by proteolytic degradation within the ER. Proteins that pass the quality check move from the ER to the Golgi apparatus via membrane vesicles. Further quality control, proteolytic cleavage or post-translational modifications occur in the Golgi. The proteins are directed to the secretory vesicles that fuse constitutively with the plasma membrane, thus releasing their contents externally.

DNA replication: the central dogma of molecular biology

Several central facts are involved in the replication of DNA and its transcription and translation into proteins (Figure 5). Firstly, DNA replication is highly 'faithful', i.e. there

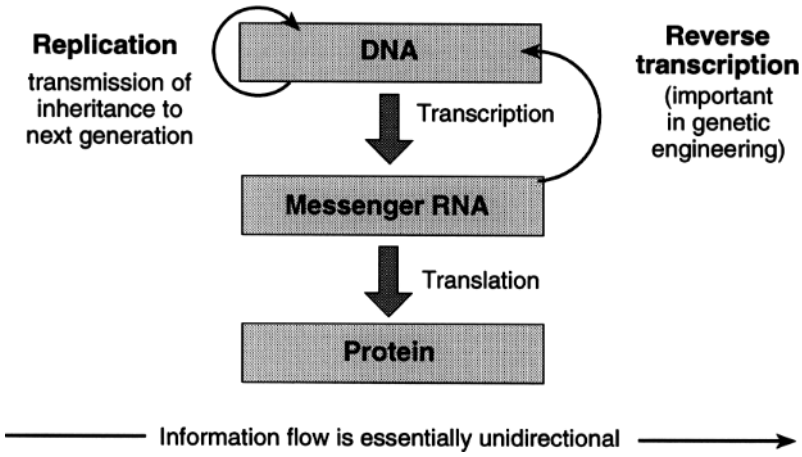

Figure 5. The central dogma of molecular genetics.

are relatively few errors (10^{-10} mutations per base pair per cell generation in the human). Second, DNA duplicates prior to cell division, which represents a DNA–DNA transfer, known as DNA replication. This is done in a semi-conservative manner (each new double-stranded daughter molecule gets one intact strand from the parent DNA, and one is newly synthesized). Third, some viruses (mainly retroviruses) use RNA rather than DNA as their genetic material. They utilize an enzyme called reverse transcriptase which can produce a single-stranded DNA molecule from single-stranded RNA. This concept has important consequences for genetic engineers. Fourthly, the flow of genetic information is always unidirectional in the mammalian cell.

DNA and the structure of the chromosome

DNA is found in the nucleus of mammalian cells in the form of a few extremely long molecules. These molecules are called chromosomes and each contains the information for constructing many different proteins. A human chromosome contains about 15% DNA, 10% RNA and 75% proteins.

In mammals, two copies (‘one pair’) of each chromosome are present in somatic cells. Each species has a typical ‘diploid’ number of chromosomes in its cells, e.g. 46 in humans. Twenty-three are inherited from the father and 23 from the mother, since each parent contributes one chromosome of each pair to the offspring. Both chromosomes of each set carry the same information, so there is a two-fold redundancy built into the system. One consequence is seen if one chromosome of the pair contains an error in a particular gene. The other chromosome, if it contains a correct copy will help remedy a mistake present in the other. If both contain the error, then the gene in question would produce an abnormal product (mutation) or be non-functional, and the physiological consequences could be life threatening.

The sole exception to this system involves the sex chromosomes. In humans, there are 22 pairs of similar chromosomes (autosomes) and in addition two sex chromosomes, either XX in the female or XY in the male. There are a number of sex-linked genetic diseases (e.g. haemophilia) which only manifest themselves in the male; this is because these diseases are linked to defective genes on the X chromosome and there is no corresponding ‘partner’ on the Y chromosome in the male.

What is a gene?

The gene can be considered as the basic unit of genetic information. In humans there are ~3 000 000 000 base pairs of DNA in the chromosomes; only a portion of this vast amount is transcribed into RNA and hence into proteins.

Only one strand of the DNA is used to encode for a particular gene. The coding strand of DNA is equivalent to the strand of mRNA produced by transcription. Thus genetic information is expressed by transcription of the non-coding strand of DNA (Figure 6).

In addition to the sequence of bases that specify the protein, there are other important regulatory sequences associated with genes: a site to ‘tell’ the cell how much RNA to make and in which tissues and under what circumstances to transcribe the DNA segment (regulatory region), a site for starting transcription and a site for stopping transcription. In mammalian cells, the story is slightly more complicated as it was discovered in the late 1970s that the genes contained extra pieces of DNA that did not appear in the mRNA that the gene encoded. These sequences are known as intervening sequences or introns. The introns have to be removed from

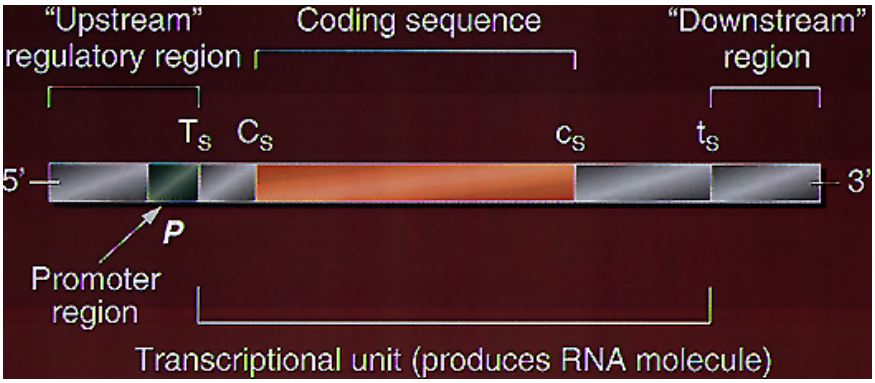

Figure 6. The gene: its organization.

the mRNA before it leaves the nucleus and is translated into a protein. For this reason, the newly synthesized mRNA is often called pre-mRNA. The DNA sequences that actually make up the fully functional mRNA (which passes into the cytoplasm) are called exons (Figure 7).

Summary

1. Proteins are large molecules made up of amino acids put together in a precise sequence.
2. The order of amino acids is determined by the sequence of base pairs in the DNA of the cell (gene).
3. DNA carries all the coded information necessary to make all of the proteins required.
4. DNA self-replicates so that the genetic information can be passed from generation to generation.
5. The DNA sequence is transcribed into mRNA.
6. mRNA sequence is translated into protein on the ribosome.

Following this basic introduction to how proteins are coded genetically and synthesized in the cell, our attention now turns to the field of genetic engineering (Section II).

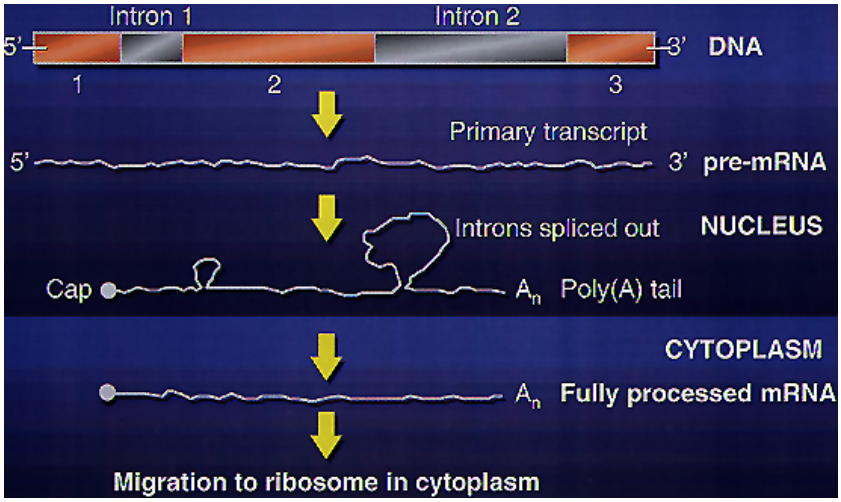

Figure 7. Structure and expression of the eukaryotic gene.

II. An introduction to genetic engineering

TABLE OF CONTENTS

Introduction

The discovery of the structure of DNA by James Watson and Francis Crick at Cambridge University, England in 1953 provided the stimulus for the development of genetics at the molecular level. Progress was rapid and by 1966 the complete genetic code had been elucidated. However, further progress on investigating the gene itself was hampered, primarily because of technical restraints. In the late 1960s, two major discoveries allowed the resurgence of research on gene manipulation. These were the discovery of an enzyme called DNA ligase (a molecular glue which joins DNA strands together) and the isolation of the first restriction enzyme (molecular scissors which cut DNA at precisely defined sequences).

Thus by 1970, the basic molecular tools (glues and scissors) were available for the construction of recombinant DNA (a DNA molecule made up of sequences that are not normally joined together, i.e. from different organisms—hybrid DNA). Figure 8 shows the production of a simple protein by recombinant DNA techniques. The process can be divided into four steps (Figure 9).

The first recombinant DNA molecules were generated at Stanford University, CA, USA, in 1972 using these basic tools. A year later this methodology was extended by joining DNA fragments to a plasmid (a small circular piece of DNA naturally found in bacteria). These recombinant molecules, when introduced into a bacterial cell, could replicate themselves. The discovery that this hybrid DNA had the ability to self-replicate had far-reaching implications and marked the emergence of the technique known as molecular or gene cloning. Today, a variety of vectors (e.g. plasmids, viruses) are employed, the choice of which depends upon the type of host cell to be used.

Uses for recombinant molecules

The recombinant molecule produced can be used for a number of purposes:

1. The ‘passenger’ DNA can be sequenced, i.e. to work out the order of the successive bases, as in the biggest undertaking to date, the Human Genome Project. It is jokingly said that this will be the most boring book ever written.
2. The function of genes can be deduced. By expressing the DNA sequence of a gene in a cell, the resulting modification can be determined or, by changing the DNA sequence of a gene, one can then determine what effect the change has on the gene’s function.
3. A gene can be introduced into an organism that is lacking or deficient in a particular function—gene therapy. This approach has been proposed to correct genetic diseases. In humans this application has raised tremendous ethical concerns.
4. A cloned DNA sequence can be used as a diagnostic tool. DNA ‘fingerprinting’ has been used in criminal cases and, more recently, this technique has been used successfully in the preimplantation diagnosis of embryos derived from parents who carry debilitating sex-linked genetic disorders. These manifest themselves primarily in male offspring (e.g. haemophilia, Tay–Sachs disease).
5. The passenger DNA can be used to synthesize a protein of interest. This is the application of most interest to us and it has led to the synthesis of the most complex proteins so far synthesized by recombinant DNA technology, the gonadotrophins follicle stimulating hormone (FSH), luteinizing hormone (LH) and human chorionic gonadotrophin (HCG).

The formation of a genomic library

The procedure of DNA cloning can be used to create a genomic library, a term describing a set of clones representing the entire genome of an organism. Here, the entire DNA of an organism can be fragmented and each fragment can be integrated into another genome, such as *Escherichia coli*, a common laboratory strain of bacteria, using an expression vector (e.g. a plasmid or phage vector) (Figure 10).

‘Recombined’ bacteria are thus obtained, each carrying a large fragment of foreign DNA. The multiplication of these bacteria then constitutes a genomic library. This contains all the constituent genes of a given DNA, including non-coding sequences (e.g. introns, control regions, repetitive sequences).

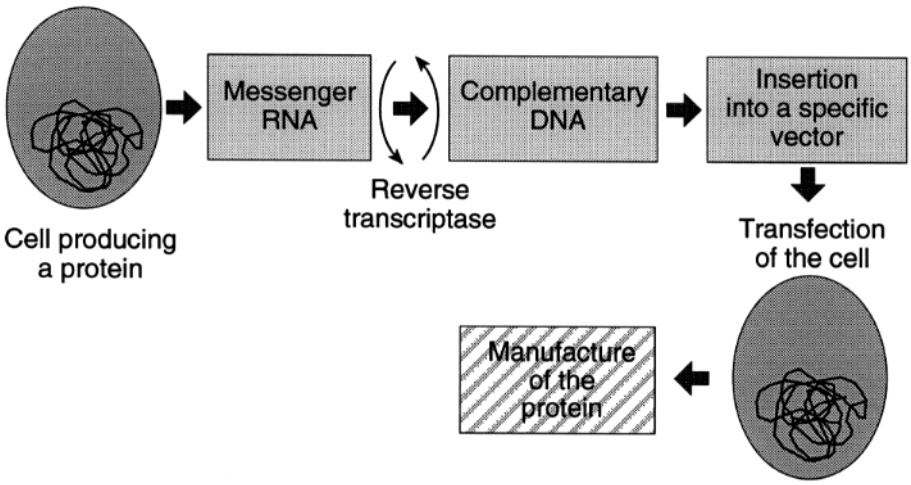

Figure 8. Production of a simple protein by recombinant DNA techniques.

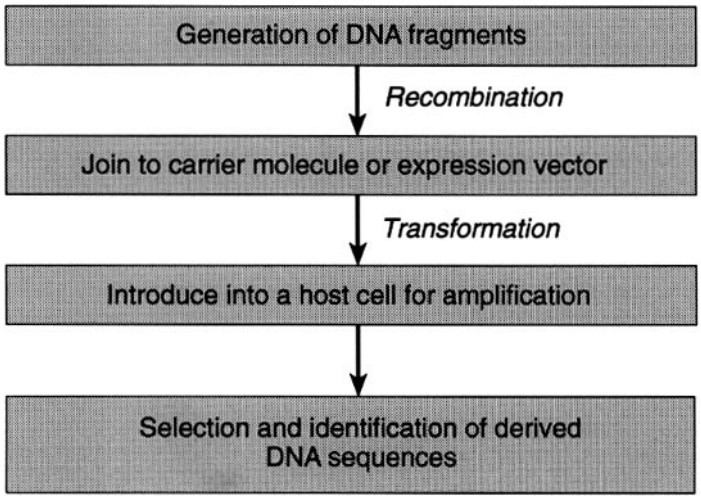

Figure 9. Genetic engineering is a four-step process.

Another, more pragmatic approach to the formation of a genomic library is the selection of functional genes, i.e. those which are expressed from the DNA. The mRNA, which is expressed from the functional genes is used in this case in order to achieve this end. The mRNA is transcribed 'backwards' by an enzyme reverse transcriptase (used by retroviruses) and eventually into a double helix DNA (called complementary or cDNA) by the action of another enzyme, DNA polymerase; then the piece of cDNA is integrated into the expression vector (plasmid) as previously described, with the aid of a ligase (Figure 11).

Cells used for genetic engineering

The host cells employed for genetic engineering belong to two categories: (i) prokaryotes, which are organisms devoid of nuclei, e.g. bacteria (*E. coli*), the DNA of which is situated in the cytoplasm; (ii) eukaryotes, e.g. superior organisms (animals, insects and plants), but also certain unicellular organisms (protozoa, yeasts, etc.), the cells of which have nuclei containing DNA organized in chromosomes.

The prokaryotes

Among the prokaryotic cells used in the laboratory (Figure 12), the bacterium *E. coli* is favoured by genetic scientists. It has particularly valuable qualities:

1. The genetic information of this bacterium is relatively modest since it codes for ~6000 different genes (in comparison with 100 000 for the superior eukaryotes), it is easy to handle and multiplies very rapidly (period for duplication is ~20 min);

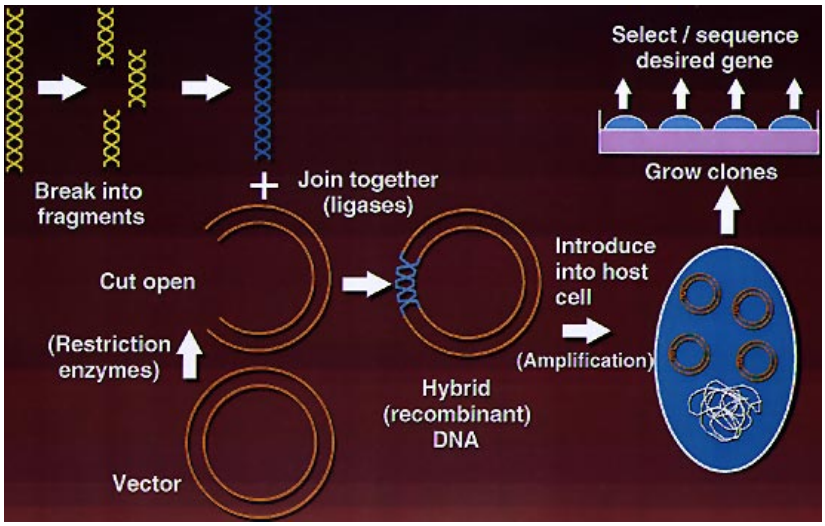

Figure 10. Gene cloning using chromosomal DNA.

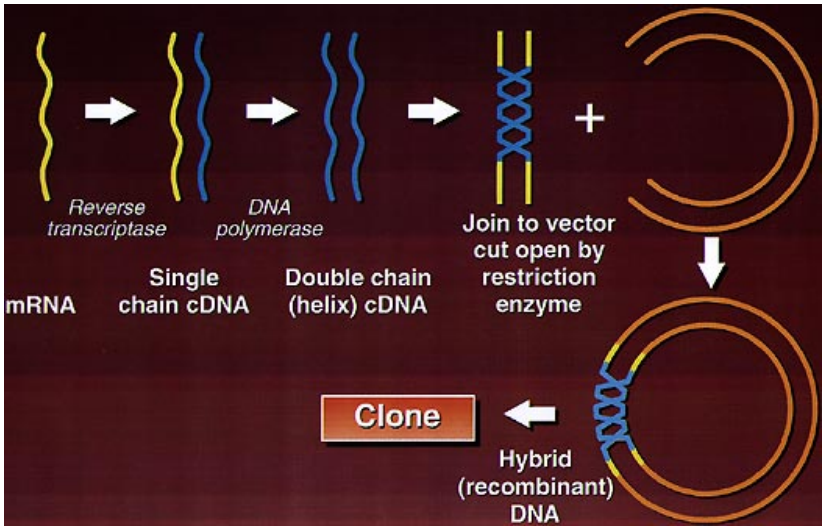

Figure 11. Gene cloning using complementary DNA (cDNA).

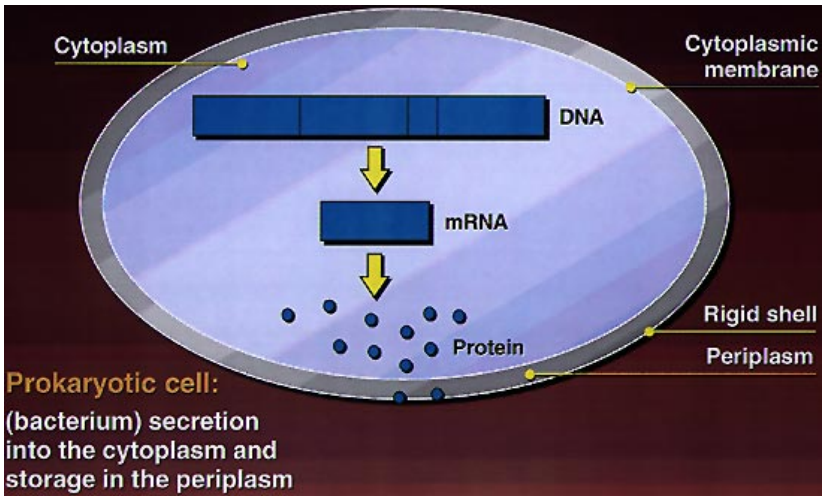

Figure 12. Protein biosynthesis in the prokaryotic cell.

2. *E. coli* may contain plasmids (self-replicating extrachromosomal fragments of circular DNA);
3. Plasmids can be introduced into *E. coli*;
4. Plasmid DNA is transcribed into mRNA which is subsequently translated at the ribosome level to obtain proteins. These stages all occur in the prokaryotic cell cytoplasm;
5. *E. coli* may produce large quantities of foreign proteins which can either be stored in the cytoplasm or 'secreted' in the periplasm.

The limits of E. coli for genetic engineering of human proteins

There are a number of limitations to the use of *E. coli* for the production of particularly complex proteins.

1. The proteins produced by *E. coli* may contain a supplementary amino acid in the N-terminal position, i.e. methionine. In the production of human proteins, it is necessary to carry out further manipulations in order to suppress this amino acid.
2. *E. coli* cannot synthesize all proteins correctly; some are wrongly synthesized and are thus without biological activity.
3. In addition, *E. coli* cannot modify the proteins after their synthesis: the process of 'maturation' is often necessary for the biological activity of the proteins (e.g. glycosylation).
4. Finally, *E. coli* stores synthesized proteins in the cytoplasm or the periplasm and does not secrete them into the culture medium. Thus supplementary manipulations are required to collect the protein, such as bacterial lysis or osmotic shock (which leads to rupture of the bacterial cell).

The advantages and disadvantages of using prokaryotic cells for genetic engineering are summarized in Table I.

Table I. Advantages and disadvantages of using prokaryotic cells for genetic engineering

Advantages	Disadvantages
Limited genetic information: 6000 genes Easy cell culture with rapid multiplication	Absence of system for maturation of proteins (e.g. glycosylation) Cytoplasmic or periplasmic storage of the proteins, necessitating further manipulation (e.g. cell rupture)
Presence of plasmids; extrachromosomal DNA easily manipulated. High levels of production/cell	Proteins produced by *E. coli* may contain a supplementary terminal amino acid, methionine*

*An enzyme system permitting splicing of methionine does not exist in the prokaryotes, which are thus incapable of transcribing an 'original' eukaryotic gene, coming to a halt at the pre-mRNA stage.

The eukaryotes

The purpose of genetic engineering in relation to human therapeutics is to obtain proteins that possess the exact characteristics of normal human proteins. It is necessary to have the specific form of the molecule and its correct folding, which requires secondary modifications and processes of maturation to replicate the protein molecule in its functional, stable and innocuous form, with activity strictly identical to that of the natural protein. Only mammalian cells offer all of these guarantees. However, using mammalian cells to engineer proteins genetically is a procedure which requires a high level of technology, because of the complexity of the molecules produced. Despite constant progress in this field, culture of eukaryotic cells is expensive and not always easy because of their complexity, and replication is slow (~24 h). In addition, similar yields per cell to those obtained using bacteria cannot be achieved. The advantages and disadvantages of using eukaryotic cells for genetic engineering are summarized in Table II.

Eukaryotic cells: a sophisticated tool for genetic engineering

As mentioned earlier, the DNA is intranuclear, being transcribed into mRNA before passing into the cytoplasm where it is subsequently translated. Further, the DNA of eukaryotic cells has, at the gene level, in addition to start and stop sequences for transcription, coding parts or exons generally interrupted by long non-coding sequences or introns.

The transcription of non-coding DNA ends in the formation of a pre-mRNA single chain carrying exons and introns present in exactly the same sequence as they appear on the coding strand of DNA. The introns, the role of which has still not been completely elucidated, are eliminated during splicing, which transforms pre-mRNA into mRNA.

The mRNA then passes into the cytoplasm (granular endoplasmic reticulum) where it is translated into protein forms (Figure 13). During their passage through the granular endoplasmic reticulum and the Golgi apparatus, the proteins take on their rigid three-dimensional structures (disulphide bridges are formed) and undergo maturation (glycosylation) before being concentrated in vesicles and finally secreted out of the cell.

In the eukaryotic cell, the gene coding for an excreted protein contains a particular sequence, the signal sequence, which is found in mature mRNA. This sequence, translated into amino acids, includes the initiation codon for mRNA. This initiation codon is composed, at the DNA level, of the triplet adenine (A), thymine (T) and guanine (G), which corresponds to the codon of adenine (A), uracil (U) and guanine (G) in the mRNA.

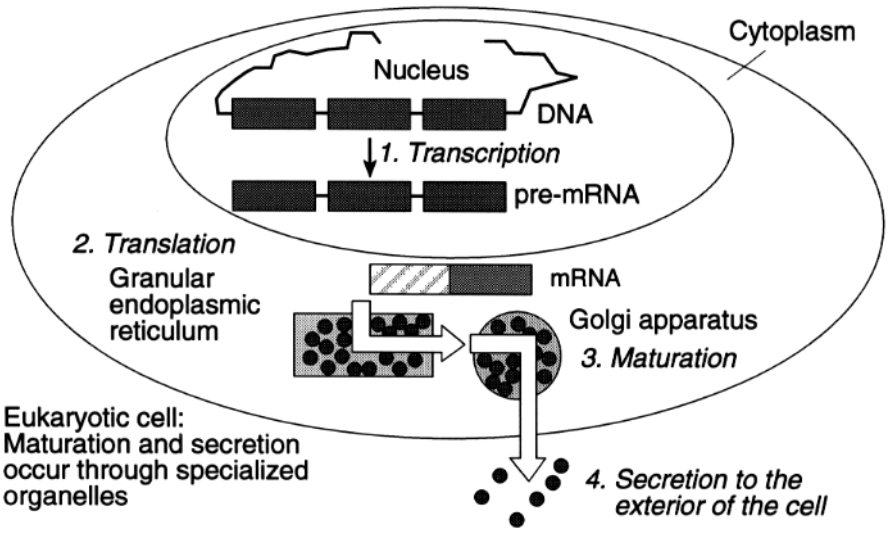

Figure 13. Protein secretion by the eukaryotic cell.

Table II. Advantages and disadvantages of using eukaryotic cells for genetic engineering

Advantages	Disadvantages
Maturation and folding of proteins assured by the ER and Golgi apparatus	Extended genetic information: 100 000 genes
Secretion of proteins into the exterior	Cell culture is difficult and expensive
Able to express complex proteins	Slow multiplication

The AUG codon corresponds to the amino acid methionine and is eliminated, with the whole of the signal sequence, after translation of the mature RNA, during the time when the protein is in the granular endoplasmic reticulum of the eukaryotic cell. As discussed previously, the bacterium does not have this specialized compartment and so is not capable of eliminating this signal sequence. Thus, the signal sequence should be deleted before expression in *E. coli*. If this is not done, then most of the proteins expressed and stored in the cytoplasm will contain an additional methionine (Met). One can avoid this by replacing the eukaryotic signal sequence with a prokaryotic signal sequence. This allows the protein to be secreted into the periplasm and the signal sequence is eliminated during the process of translocating the protein from the cytoplasm to the periplasm.

The methods of production of proteins by prokaryotic and eukaryotic cells therefore show fundamental differences (Table III).

In order to express and synthesize such complex proteins as the gonadotrophins it was necessary for the Serono scientists to use the mammalian cell.

Table III. *Escherichia coli* versus mammalian cells: post-translational events

Post-translational events	Cell type		
	E. coli	Mammalian	r-hFSH
Protein maturation	yes	yes	✓
Secretion	no	yes	✓
Methionine removal	no/yes	yes	✓
Disulphide bridge	yes	yes	✓
Glycosylation	no	yes	✓
Phosphorylation	no/yes	yes	
Amidation	no	yes	
Myristoylation	no	yes	
Carboxylation	no	yes	
Heterodimerization	yes	yes	

Summary

1. Genetic engineering is a technology allowing the manipulation of DNA/genes in various ways to achieve certain goals in both pure and applied science and medicine.
2. Vectors are DNA molecules able to replicate autonomously in a host cell.
3. Prokaryotic or eukaryotic cells can be used as the host cell.
4. Eukaryotic cells are the hosts of choice for the production of complex molecules.

In the next section, the development of a recombinant form of the human gonadotrophin FSH, in relation to its expression in a host mammalian cell, and its characterization and commercial production are described.

III. Expression of human FSH (Gonal-F®) by recombinant DNA technology

TABLE OF CONTENTS

What is human FSH?

Human FSH (hFSH) is a dimeric glycoprotein (i.e. composed of two protein subunits) to which there are attached four complex carbohydrate structures. Its composition is as follows: It is a glycoprotein hormone consisting of two non-covalently linked, non-identical protein components (α- and β-subunits; Table IV). The α-subunit is composed of 92 amino acids and carries two carbohydrate moieties linked to Asn-52 and Asn-78; the β-subunit is composed of 111 amino acids and carries two carbohydrate moieties linked to Asn-7 and Asn-24. The FSH molecule exists in many different (iso)forms; the microheterogeneity is due to differences in carbohydrate moieties (Table V). The carbohydrate moieties are important in determining the half-life of FSH and therefore modulate in-vivo biological action.

The FSH present within the anterior pituitary gland exists as a heterogeneous population, i.e. a family of isoforms (structural variants of a given protein). The biochemical basis of this heterogeneity lies not in differences in the primary structure of the molecule but rather in the post-translational modifications that occur at the site of the granular endoplasmic reticulum and the Golgi apparatus in the mammalian cell.

The FSH isoforms are identical in terms of amino acid sequence of the two peptide subunits (α and β) and the attachment points of the carbohydrate side chains (four in total, two on the α and two on the β-subunits). However, what does vary is the composition of the carbohydrate side chains themselves. The chains can exist in many branched forms which may or may not be 'capped' by sialic acid residues.

These multiple forms of FSH differ in their plasma half-life and therefore in their biological activity. In a similar way to other glycoprotein molecules, FSH is removed from circulation by binding to the asialoglycoprotein receptor on the plasma membrane of the liver and kidneys. FSH isoforms with few sialic acid residues are quickly removed from circulation following binding to the liver and kidney receptor. Those isoforms which are heavily sialated escape capture by the receptor and therefore reside for longer periods in the circulation and have greater in-vivo bioactivity. Thus, more acidic FSH isoforms (those with more sialic acid residue caps) have a longer half-life (residence time in the blood) than less acidic (more basic) isoforms (those with fewer sialic acid residues) (Table VI). Menopausal urine contains a majority of the more acidic isoforms.

New research suggests that the carbohydrate structures also play a major role in determining the biological activity of FSH compared to that of other glycoprotein hormones. Experiments have shown that certain of the carbohydrate structures (two on the β-subunit) play an essential role in determining the plasma halflife and thus the in-vivo bioactivity of FSH.

Expression of hFSH by mammalian cells

Introduction

As discussed previously (Section II), the 'work horse' of genetic engineering, *Escherichia coli* (a prokaryotic cell), does not have the intracellular machinery (endoplasmic reticulum and Golgi apparatus) necessary to fold complex proteins correctly and add carbohydrate structures to proteins. As these are fundamental to the biopotency of the FSH molecule, it was necessary for the genetic engineers to turn to the mammalian cell for assistance in producing biologically active hFSH. The expression and production on a commercial scale of r-hFSH has been a huge achievement as FSH is one of the most complex proteins to date produced by recombinant DNA (rDNA) technology.

Below is a brief summary of the background of how hFSH was expressed in a mammalian cell system by rDNA technology. More information is given in Chappel *et al.* (1992)[1].

[1]Chappel, S.C. *et al.* (1992) Expression of human gonadotropins by rDNA methods. In Genazzani, A.R. and Pettraglia, F. (eds), *Humans in Gynecological Endocrinology.* Parthenon Press, Carnforth, UK, pp. 179–184.

Table IV. Characteristics of human FSH

Subunit	Molecular size (kDa)	No. of carbohydrate structures	Disulphide bridges	Properties
α	14	2	5	Subunit common to all the gonadotrophins
β	17	2	6	Confers biological and immunological specificity

Table V. Microheterogeneity of FSH (isoforms)

Biochemical basis due to post-translation modifications
Specific differences in carbohydrate moieties added to protein backbone and in terminal sialic acid residues
These multiple forms differ in their plasma half-life and hence in their biological activity
Less acidic isoforms are removed more quickly from the circulation and therefore have lower potency *in vivo*
More acidic isoforms reside longer in circulation and therefore have higher potency *in vivo*

Table VI. FSH isoforms

Type of FSH isoform	Sialic acid content	Half-life *in vivo*
Acidic	High	Long
⇕	⇕	⇕
Basic (less acidic)	Low	Short

Early attempts to genetically engineer hFSH were carried out by inserting the relevant genetic information (cDNA for the α- and β-FSH genes obtained from pituitary cDNA libraries) into separate expression vectors which were transfected into a mouse fibroblast cell line. Biologically active FSH was expressed by this method but the yield was low, so, in order to improve the expression system, the scientists went back to the drawing board and re-evaluated how hFSH is made *in vivo*. First it was necessary to isolate the complete human α- and β-FSH genes.

Isolation of the α- and β-hFSH genes

Preparation of the human genomic library

Human DNA was isolated and fragments were attached to a vector (bacteriophage) and the recombined bacteriophage was used to infect *E. coli*. The recombinant bacteriophages were amplified so as to establish a permanent library of cloned human DNA fragments.

Isolation of the α-hFSH gene

This was done by using the sequence information from the cDNA clone of α-HCG (the glycoprotein hormones are a group of structurally related molecules which share a common α-subunit) to construct suitable probes for the identification of a full-length α-hFSH genomic clone. A suitable probe was used to select a particular clone present in the human genomic library. The α-subunit gene is composed of four exons interrupted by three intron sequences, and ~700 nucleotides encode for the mature transcript.

Isolation of the β-hFSH gene

The human genomic library described above was screened by using two probes designed from the partially known amino acid sequence of the β-hFSH. Screening the 'library' with one of the two probes led to the isolation of one clone. The β-hFSH gene consists of three exons with two intervening introns. Following cloning from the human genomic library, the full-length α- and β-hFSH subunit genes were confirmed by DNA sequencing and then placed into mammalian cell vectors designed for high expression of the subunit mRNAs.

Incorporation of the α- and β-hFSH genes into a host cell

It is well established that within the pituitary gonadotroph, the α-subunit protein is present in excess compared to the β-subunit. Additionally, there are a number of external factors (e.g. gonadal steroids, gonadotrophin-releasing hormone) which have been shown to regulate the transcription of gonadotrophin mRNA and also the rate of degradation of mRNA. It has in addition been shown that the presence of introns can influence mRNA expression. In the intron region the existence of mRNA enhancer elements has been demonstrated. Thus, from a molecular biology viewpoint, it was decided to incorporate each subunit gene (complete with introns) into a separate suitably designed expression vector which also contained separate amplifiers (genes whose expression rates can be controlled by altering external conditions). For example, the α-gene was linked to dihydrofolate reductase (DHFR), which is required for the synthesis of nucleic acid precursors.

The α and β expression vectors were co-transfected into another very well characterized mammalian cell line, the Chinese hamster ovary (CHO) cell line (Figures 14 and 15). CHO cells are extensively used in genetic engineering to produce other marketed complex proteins such as erythropoietin. In this case, the particular CHO cell line used was DHFR deficient. The central role of DHFR in the synthesis of nucleic acid precursors together with the sensitivity of DHFR-deficient cells to tetrahydrofolate

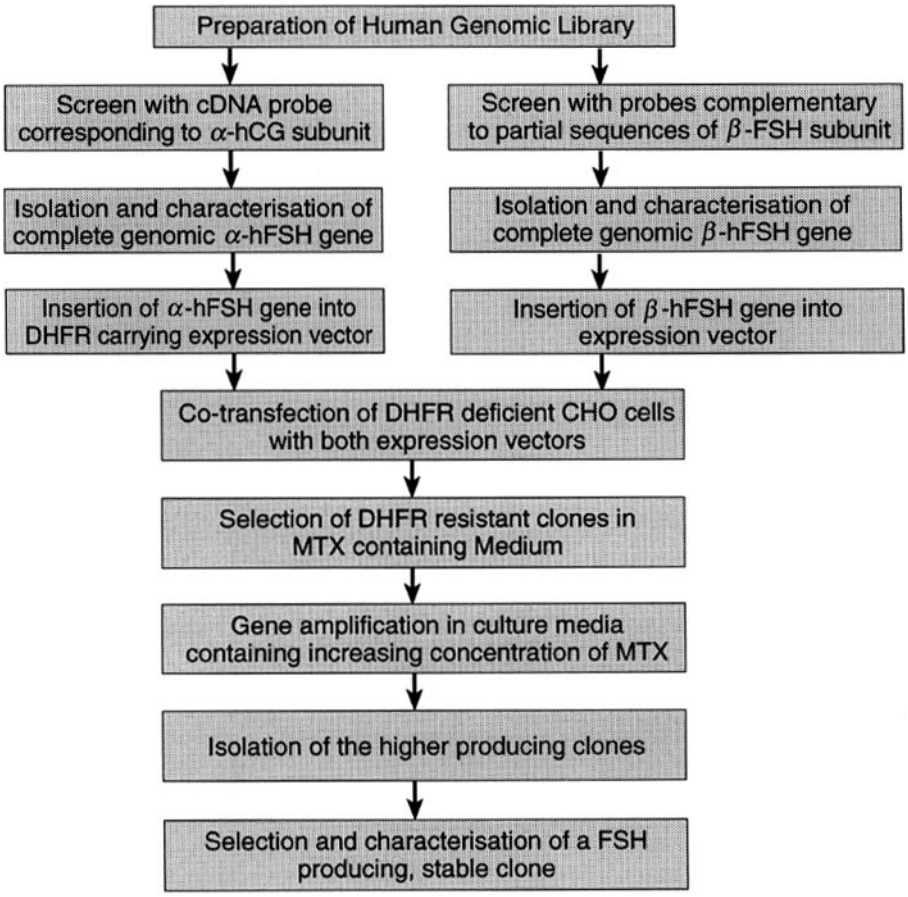

Figure 14. Expression of hFSH in CHO cells.

analogues such as methotrexate (MTX) offer two major advantages for genetic engineering. Firstly, transfection of such DHFR-deficient cells by vectors containing a DHFR gene allows the selection of recombinant MTX-resistant cells when cultivated in media containing MTX. Secondly, culture of these cells in selective media containing increasing concentrations of MTX results in amplification of the DHFR gene and the associated DNA.

Following transfection with the two expression vectors carrying the α- and β-hFSH genes, the transfected cells were cultured in conditions where the concentration of MTX was increased in a stepwise fashion. This led to a co-amplification of the - and β-hFSH genes (Figures 14 and 15).

The establishment of the master cell bank

Following transfection of the CHO cells, a group of transformants was isolated that were genetically stable and had adequate levels of productivity of biologically active r-hFSH. This was followed by a very detailed evaluation of one particular cell line which displayed promising characteristics (high and stable level of FSH expression). Following a series of cell expansion and evaluation steps, one particular CHO cell transfectant was chosen on the basis of cell line productivity, stability over time and quality of the molecules secreted. This was used to establish the master cell bank (MCB). The MCB consists of individual vials containing identical cells. These are cryopreserved until required. This clonal line has been fully characterized in terms of its genetic stability and structure.

Table VII. Summary of the establishment of the master cell bank (MCB) and working cell bank (WCB)

Following transfection, individual clones (derived from a single cell) were screened on the following basis:
Cell (FSH) productivity
Biological activity of FSH
Genetic stability
One CHO cell transfectant was selected
MCB was established from cells of this single selected clone
MCB vials were fully tested and cryopreserved until required
WCB was established by expansion of cells from a single MCB vial
One or more WCB vials were used for a production run

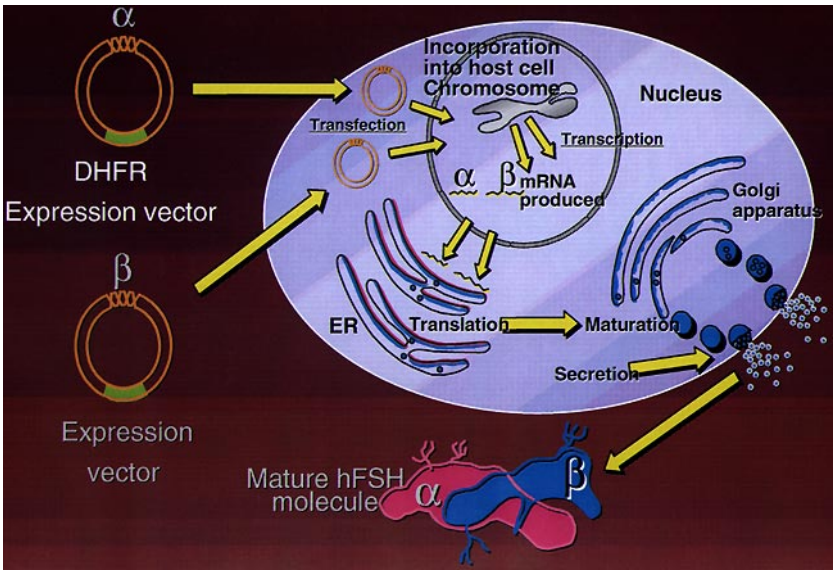

Figure 15. Expression of hFSH in CHO cells.

A working cell bank (WCB) was then established by expansion of cells recovered from a single vial of the MCB (Table VII). The cells were successively expanded and then aliquots were put into vials and cryopreserved. One or more of the WCB vials is used for each production run. Both the MCB and WCB have been tested for sterility, mycoplasma and viral contaminants and other adventitious agents according to EEC and FDA guidelines. They have also been tested for genetic stability and structure.

Analysis has shown that both genes are incorporated into closely related chromosomal locations in the genome of the recombinant CHO cell.

Commercial production of recombinant hFSH (Gonal-F®)

The production process for bulk r-hFSH (see Figure 16) from genetically engineered CHO cells consists of two major stages: (i) a cell culture process, resulting in r-hFSH present in a crude liquid supernatant, and (ii) a downstream process for the purification of r-hFSH from the crude liquid supernatant.

The cell culture process for the production of r-hFSH is based on the large-scale culture of the production clone in a bioreactor. WCB vials containing the CHO cell line transfected with the human genes encoding for the α- and β-FSH subunits are grown in a bioreactor equipped to allow precise control of the culture parameters (temperature, pH, dissolved oxygen, etc.) and the nutrient composition of the medium.

The culture process consists of a scaling-up phase (expansion of cells) and a r-hFSH production phase characterized by a continuous perfusion of fresh medium (Figure 16). Cells are multiplied up initially from one vial of the manufacturer's WCB in flasks and then in roller bottles until a sufficient cell number for inoculation of the bioreactor is achieved. The cells are then mixed with a suspension of microcarrier beads (on to which the cells attach) and transferred to the bioreactor vessel. The reactor is perfused with a medium selected to promote cell attachment and growth. After a defined period, the perfusion medium is gradually changed to a medium suitable for the production phase of the culture. The r-hFSH production phase then follows in which conditioned medium is collected for downstream purification. (Tests were also done on early production runs, where cells were taken and re-analysed to ensure that they had not undergone any degenerative changes.)

The purification of r-hFSH from harvested culture supernatants is divided into six steps: an ultrafiltration step and five chromatographic steps (Figure 17). The principal purification is achieved through an immunoaffinity chromatographic step (using a murine-derived anti-FSH monoclonal antibody, step III in Figure 17). This step is similar to that used in the manufacture of highly purified urinary FSH (Metrodin HP). The remaining steps in the purification process increase the purity of the final r-hFSH bulk product. Each step of the production is rigorously controlled so as to ensure the consistency of the final product.

Product definition

Gonal-F® is the proprietary name of the final product and the synonym is recombinant human follicle stimulating hor-

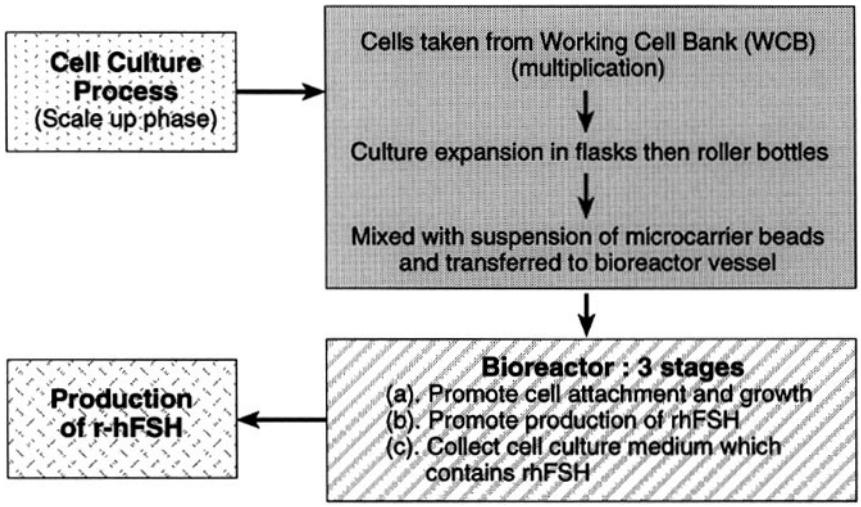

Figure 16. Flow chart summarizing the process for bulk production of r-hFSH.

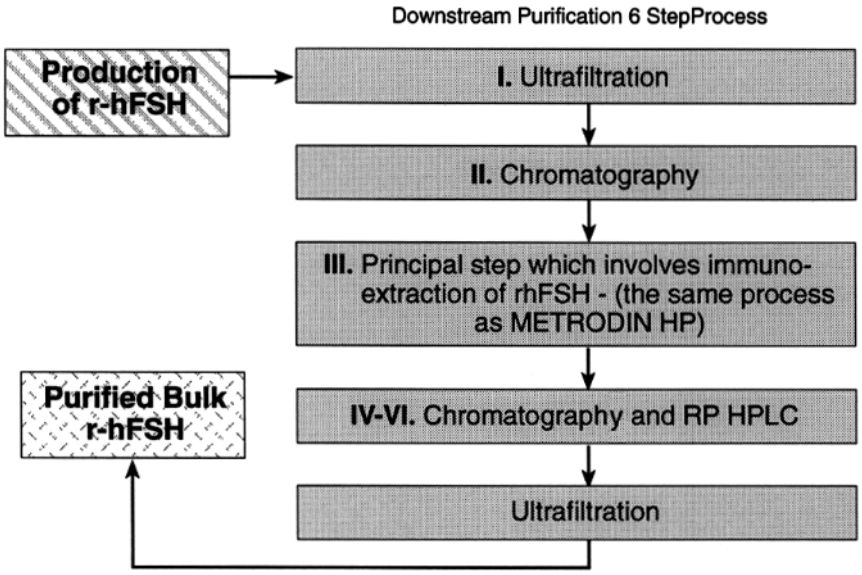

Figure 17. Flow chart summarizing production of r-hFSH from harvested culture supernatants.

mone (r-hFSH). It is made by Ares-Serono, Switzerland. A similar product, Puregon, is made by Organon, The Netherlands.

Physicochemical characteristics

Gonal-F® r-hFSH consists of two non-covalently linked, non-identical protein components designated as the α- and β-subunits. The α-subunit is composed of 92 amino acids carrying two carbohydrate complexes linked to amino acid Asn-52 (asparagine) and Asn-78. The β-subunit is composed of 111 amino acids carrying two carbohydrate complexes linked to Asn-7 and Asn-24.

The full amino acid sequences of the α- and β-subunits of the r-hFSH, as determined by DNA sequencing of the cDNA and by directly sequencing the protein subunits, were found to be identical to natural (urinary and pituitary) hFSH (Figure 18). On the basis of laser densitometric mass spectrometry, the relative molecular masses of the α- and β-subunits have been determined to be ~14 and 17 kDa respectively.

Isoforms of r-hFSH

Amino acid sequencing has determined localization of the glycosylation sites at α-Asn-52, α-Asn-78, β-Asn-7 and β-Asn-24, which are identical to those observed for native

a-**subunit:**

Ala Pro Asp Val Gln Asp Cys Pro Glu Cys	10
Thr Leu Gln Glu Asn Pro Phe Phe Ser Gln	20
Pro Gly Ala Pro Ile Leu Gln Cys Met Gly	30
Cys Cys Phe Ser Arg Ala Tyr Pro Thr Pro	40
Leu Arg Ser Lys Lys Thr Met Leu Val Gln	50
Lys **Asn** Val Thr Ser Glu Ser Thr Cys Cys	60
Val Ala Lys Ser Tyr Asn Arg Val Thr Val	70
Met Gly Gly Phe Lys Val Glu **Asn** His Thr	80
Ala Cys His Cys Ser Thr Cys Tyr Tyr His	90
Lys Ser	92

b-**subunit:**

Asn Ser Cys Glu Leu Thr **Asn** Ile Thr Ile	10
Ala Ile Glu Lys Glu Glu Cys Arg Phe Cys	20
Ile Ser Ile **Asn** Thr Thr Trp Cys Ala Gly	30
Tyr Cys Tyr Thr Arg Asp Leu Val Tyr Lys	40
Asp Pro Ala Arg Pro Lys Ile Gln Lys Thr	50
Cys Thr Pro Lys Glu Leu Val Tyr Glu Thr	60
Val Arg Val Pro Gly Cys Ala His His Ala	70
Asp Ser Leu Tyr Thr Tyr Pro Val Ala Thr	80
Gln Cys His Cys Gly Lys Cys Asp Ser Asp	90
Ser Thr Asp Cys Thr Val Arg Gly Leu Gly	100
Pro Ser Tyr Cys Ser Phe Gly Glu Met Lys	110
Glu	111

Figure 18. Amino acid sequences of r-hFSH (GONAL-F®). **Asn** = *N*-glycosylation sites.

hFSH. The structures of the carbohydrate moieties of r-hFSH have been elucidated from mass spectrometry studies and found to be representative of forms found in pituitary hFSH. A list of the tests carried out on the recombinant generations of products and an old generation gonadotrophin (Metrodin) is given in Table VIII.

The isoforms can be separated to a certain degree, according to their relative charge (isoelectric point, p*I*) by isoelectric focusing (IEF). The separated protein bands can be visualized by staining them (commonly with Coomassie brilliant blue). However, in order to confirm that these protein bands are actually FSH isoforms, a more specific technique is used called 'Western Blotting'. In this technique, monoclonal antibodies to FSH are used which 'tag' the FSH isoform molecule if it is present. This is a highly specific tagging technique and is routinely used to monitor the isoforms present in each batch of r-hFSH.

Preparations of r-hFSH demonstrate a highly consistent IEF profile which lies within the p*I* range 3.5–6.1. This compares with an IEF range of p*I* 3–5.5 for Metrodin HP. Thus, Gonal-F® contains slightly more basic FSH isoforms than Metrodin HP. The physicochemical characteristics of Metrodin, Metrodin HP and Gonal-F® are shown in Table IX.

Regardless of the inherent complexity of the protein, batches of bulk r-hFSH show consistency with respect to physicochemical characteristics and immunological and biological activity. With regard to the biological activity of r-hFSH, which is measured according to the rat ovarian weight gain assay, the recombinant derived molecule is indistinguishable from hFSH extracted from urine.

Table VIII. Tests used to evaluate FSH preparations

Product specification	Measuring tools		
	Old preparation	Metrodin HP	Gonal-F®
Activity/potency	In-vivo bioassay	In-vivo bioassay	In-vivo bioassay
Identity	In-vivo bioassay	IEF	IEF
		Peptide mapping	Peptide mapping
		Subunit sequencing	Subunit sequencing
			Oligosaccharide structure
Composition	Not done	Amino acid analysis	Amino acid analysis
		Carbohydrate analysis	Carbohydrate analysis
		Terminal sugars	Terminal sugars
		Sialic acid	Sialic acid
Carbohydrate (%)	Not done	HPAE	HPAE
		Chromatography	Chromatography
Purity	Not done	SE-HPLC	SE-HPLC
		RP-HPLC	RP-HPLC
		SDS-PAGE silver stain	SDS-PAGE silver stain
			Cell culture-derived proteins

Table IX. Physicochemical analysis and product release specifications of urinary and recombinant gonadotrophin preparations

Specification	Metrodin	Metrodin HP	Gonal-F®
Potency	In-vivo bioassay	In-vivo bioassay	In-vivo bioassay
Specific activity (FSH/mg protein)	≡ 150 IU	≡ 10 000 IU	≡ 10 000 IU
Protein content/75 IU (μg)	370–750	6–11	6–11
Active protein content (% FSH) in bulk	<3%	>95%	≥97%
Residual LH activity	0.7 IU per 75 IU FSH	Negligible	None
IEF range	?	3–5.5	3.5–6.1

Summary

1. r-hFSH is expressed in mammalian (CHO) cells.
2. The starting material (CHO cells) has been extensively characterized.
3. There is a two-phase production process for r-hFSH (Gonal-F®): cell culture and purification.
4. The cells used are derived from a single cell line in all production runs.
5. r-hFSH has greater purity than urinary hFSH.
6. r-hFSH has a consistent isoform profile.
7. There is no contamination by other human proteins such as human luteinizing hormone.
8. Standardized procedures can be used for bulk production.

Milestones in biotechnology and the development of recombinant human FSH (Gonal-F®)

1972	First recombinant DNA molecules generated at Stanford University, CA, USA.
1973	Foreign DNA joined to plasmid was able to self-replicate when introduced into a bacterial cell (first example of cloning).
1977	The hormone somatostatin (14 amino acids long) was expressed from recombinant DNA.
1979	Human insulin was cloned in bacterial cells.
1979	Human chorionic gonadotrophin α-subunit cloned.
1979	Human growth hormone was cloned in bacterial cells.
1982	Genetically engineered human insulin approved in the USA and UK for use in the treatment of diabetes.
1983	The correct amino acid sequence of the human β-FSH subunit was established.
1985	Human β-FSH gene was cloned and biologically active FSH was expressed in mouse fibroblast cells.
1987	Using a segment of the human β-FSH gene as a probe, the β-FSH gene was found to be localized to human chromosome 11.
1988	Human FSH was successfully expressed in a CHO cell.
1989–90	Serono established the Master Cell Bank (MCB) and Working Cell Bank (WCB) from which all cells producing r-hFSH originated.
1990	First batches of r-hFSH for clinical use were made in Switzerland.
Mar. 1991	r-hFSH was first used in the clinic.
Nov. 1991	Start of phase III studies using r-hFSH.
May 1992	Report in the *Lancet* of the first pregnancies following the use of r-hFSH in women undergoing in-vitro fertilization in Switzerland and Belgium.
Sept. 1993	Registration files submitted in Europe and USA.
Aug. 1994	First registration worldwide for the clinical use of Gonal-F®.
Oct. 1995	Gonal-F® was the first pharmaceutical to be granted European-wide marketing approval by the European Medicines Evaluation Agency.

GLOSSARY

Amino acids The basic building blocks of proteins
Amplification Duplication of a DNA sequence
Anticodon Three bases on a tRNA molecule that are complementary to the codon on mRNA
Bacteriophage A bacterial virus (phage)
cDNA Complementary DNA. A copy of DNA obtained from a mRNA by the action of reverse transcriptase
cDNA library A collection of clones prepared from the mRNA of a given cell or tissue type, representing most of the genetic information expressed by such cells
Chromosome A DNA molecule carrying a set of genes. Bacteria have a single chromosome, eukaryotic cells have multiple chromosomes
Clone A set of genetically identical cells or organisms derived from a single individual
Codon Decoding unit for the genetic code, each codon is composed of three nucleotides; the assembly of codons makes up the mRNA molecule
DNA Deoxyribonucleic acid. Basic material of heredity, constituting the chemical basis of the chromosomes
DNA polymerase The enzyme which is capable of synthesizing double-stranded DNA by following instructions from a single-strand DNA
Disulphide bridge A covalent linkage between amino acids (cysteines) at different points in a protein. These are important for protein structure
Enhancer A sequence of DNA which enhances, indirectly, transcription starting from the promoter sequence of a gene. May be remote from the promoter
Eukaryote An organism whose cellular nucleus is enclosed in an envelope: superior organisms (animal, plant), unicellular organisms (yeast, protozoan)
Exon Coding part of the DNA from eukaryotic genes
Expression vector Vector which, when introduced into a host cell, may induce the expression of a target gene
Gene A unit of genetic information carried by the chromosomes; one gene ≡ one protein
Genetic code Cellular dictionary establishing the correlation between the codons (base triplets) and the amino acids that constitute proteins
Genome The total genetic information of an organism
Genomic library A collection of clones which together represent most of the genome of an organism. Amplification is normally carried out in bacteria (*Escherichia coli*) using an expression vector
Glycoprotein A protein which carries carbohydrate moieties
Glycosylation A final processing stage of proteins, essential for the function of some, which can only be carried out by eukaryotic cells
Intron Intervening sequences: non-coding part of the DNA present in most eukaryotic genes
Isoelectric focusing A technique used to separate proteins according to their charge
Isoform A structural variant of a protein
Ligase Enzyme permitting the molecular joining together of two fragments of DNA (= molecular glue)
Master cell bank (MCB) Consists of a series of vials containing a given number of MCB cells derived from a single recombinant cell
mRNA Messenger RNA which is a single-stranded chain. Mature RNA capable of being decoded by the ribosomes and translated into proteins
Nucleotide Basic element of DNA; each nucleotide is composed of one of four bases (guanine, thymine, cytosine or adenosine), a sugar and a phosphoric acid
Plasmid Self-replicating extra-chromosomal fragment of DNA of circular form naturally present in bacteria. Plasmids are the carriers of genes responsible for specific properties (e.g. bacterial resistance to antibiotics)
Pre-mRNA The RNA sequence transcribed from an eukaryote gene which consists of both exons and introns. The introns are spliced out prior to the mRNA leaving the nucleus
Probe A piece of DNA (usually labelled with radioactivity) that recognizes and binds specifically to a complementary sequence in another sample of DNA
Prokaryote Simple unicellular organism devoid of a nuclear envelope, e.g. bacterium

Promoter sequence Specific gene sequence which permits the initiation and effective transcription (DNA passage to RNA) of the gene to which it is attached
Recombined bacterium Bacterium which has been modified by introducing foreign DNA
Regulator sequence Specific gene sequence which permits control of the level of expression of the gene
Restriction enzymes Enzyme capable of cutting the two chains of DNA at a highly specific position (≅ molecular scissors)
Reverse transcriptase Enzyme normally found in viruses which is capable of synthesizing DNA from mRNA
RNA Ribonucleic acid. Different from DNA in that it contains ribose sugars and the base uracil (U) instead of thymine (T). In most instances RNA is single stranded
Splicing Elimination of non-coding sequence (intron) of pre-mRNA, permitting the formation of mature mRNA
Transcription Synthesis of mRNA from DNA
Transfection Integration of a fragment of foreign DNA into eukaryotic cells
Transformation Integration of a fragment of foreign DNA (usually plasmid) into prokaryotic cells
Transformant A cell which has been modified due to the integration of new DNA
Translation Synthesis of protein by the ribosomes from the mRNA
Vector A molecule of DNA which can be introduced into a host cell and incorporate foreign DNA into the genome of that cell or to replicate therein as an autonomous unit. They can be plasmids, phage particles, etc.
Western blotting A technique for analysing electrophoresed proteins which have been transferred to filter paper. These proteins can be tagged by specific antibodies
Working cell bank (WCB) Expansion of cells recovered from a single vial of the MCB. These are used in the production runs.