Biology Today

All in the Family

Maura C. Flannery
Department Editor

Families are on my mind right now. It's a week before Christmas as I write this column, and I'm looking forward to having the house full of family over the holidays. So it seems appropriate to write about families of genes: groups of genes that have similar nucleotide sequences over at least parts of their length. The alpha and beta chains of hemoglobin are examples of related proteins coded by related nucleotide sequences. The immunoglobulins are obviously examples as well. But sequence similarities do not necessarily relate to functional similarities. N-CAM, a protein involved in the adhesion of brain cells, bears a family resemblance to immunoglobulins, as do some muscle proteins (Benian et al. 1989). It is these surprising relationships that make genetic genealogy so interesting. The tools now available to molecular biologists allow them to work out nucleotide sequences more quickly than in the past. That's why the number of families and family members is growing so rapidly. Sophisticated computer programs make it possible to compare newly-found sequences with those already entered into data banks. This is how similarities were discovered between such different proteins as lens crystallin and alcohol dehydrogenase.

Maura C. Flannery is associate professor of biology and chair of the division of computer science, mathematics and science at **St. John's University, Jamaica, NY 11439**. She earned a B.S. in biology from Marymount Manhattan College, an M.S., also in biology, from Boston College and a Ph.D. in science education from New York University. Her major interest is in communicating science to the nonscientist. She has developed biology courses for criminal justice majors and for communications and journalism majors, as well as courses in reproductive biology and exercise physiology.

Among the most crucial questions to be asked about gene families are why do such families exist, and how did they develop? The answers, in most cases, seem to involve gene duplication (Maeda & Smithies 1986). During meiosis, if crossing over is nonhomologous, not reciprocal, two copies of a gene can end up on one chromosome. Since only one gene is needed, mutations in the other may not be selected against. Such mutations usually make the gene inactive or the protein product useless, but in rare cases the protein can have new and useful properties. Then this new genetic sequence would be selected for in future generations. Another way genes can be duplicated is through amplification, which happens during DNA replication if initiation of replication occurs more than once at the same point. Amplification has been found frequently in somatic cells, but it is still uncertain if it occurs in meiosis, so it may or may not be important in the evolution of gene families.

Still another way genes might be duplicated is through an RNA intermediate. A reverse transcriptase enzyme could allow an RNA sequence to be used as a template for DNA, and the resulting DNA sequence could be inserted into a chromosome. Evidence for this process comes from genetic sequences that have strings of adenine at their 3' ends, as messenger RNA does. These genes also lack the introns or intervening sequences that are found in most eukaryotic genes. These are DNA sequences that are transcribed into RNA, but do not code for amino acids. Introns are removed from, spliced out of, the RNA messenger before translation into protein. If such a spliced messenger were then transcribed back into DNA, the resulting gene would lack intervening sequences. Because such genes are formed from RNA transcripts, the copies usually don't have regulatory sequences and thus are inactive. But, by chance, they may pick up regulatory sequences at the insertion site.

Nuclear Receptor Family

With this brief review of how families develop in mind, we can now look more closely at one particular family. There are a large number to choose from, but I think I'll start with a "superfamily." The term does not designate the genes of Superman, but rather designates a family of families. Some of the genes in a superfamily might be thought of as siblings and others as distant cousins. One of the most diverse such groups is the steroid and thyroid hormone receptor superfamily (Evans 1988). These intracellular receptors, after binding hormones, then bind to DNA and trigger transcription of specific genes. Belonging to this family are receptors for all three classes of steroid hormones: the adrenal steroids (e.g. cortisol, aldosterone), the sex hormones (e.g. estrogen, progesterone, testosterone) and vitamin D_3. But the family is still larger. It includes the receptors for thyroid hormone and for retinoic acid, an important vertebrate morphogen. So this nuclear receptor family is crucial in embryology as well as in physiology.

But the molecular genealogists aren't finished yet! They have found still other relatives in this superfamily: the receptors for phenobarbital and for dioxin. These receptors do not appear to bind to any known naturally occurring chemical in the body, but why should animals have high-affinity receptors for such foreign substances? Daniel Nebert (1990) notes that:

> Clearly, the organism has not been waiting hundreds of millions of years to be challenged by drugs or environmental pollutants. I suggest that these foreign chemicals are structurally similar enough to mimic an endogenous

ligand that has an important role in growth or morphogenesis.

The search for this ligand may yield to still other family secrets and information on how such foreign molecules cause problems in the body. In fact, the family members themselves can also cause problems. The cancer-causing gene or oncogene, v-*erb* A, is related to the gene for the thyroid hormone receptor. v-*erb* A produces a protein involved in the development of a form of leukemia. Without binding thyroxin as the normal receptor would, this protein can bind to DNA and trigger transcription. In the future, many such family relationships between normal and abnormal genes will no doubt be found. The proteins of the GTPase family, discussed below, are related to the RAS proteins involved in several types of cancer.

Protease Inhibitors

Knowing about gene families can not only help in the understanding of disease, but in its treatment as well. The protein α_1-antitrypsin is a natural inhibitor of protease enzymes and belongs to the family of serine protease inhibitors, the serpins. The members of this family are all globular proteins; they also share the same protein-cutting function. In each, the reactive center is a loop on the exterior of the molecule. The proteins differ in the amino acid residue at what is called the P_2 site on the loop. In α_1-antitrypsin, whose specific target is elastase, the P_1 amino acid is methionine. In a mutant form of α_1-antitrypsin found in humans, the methionine is replaced by valine and the protein is inactive. Individuals homozygous for this defect are susceptible to emphysema because, since the elastase is not inhibited, elastic tissue in the lungs is damaged (Carrell 1984). Also, normal α_1-antitrypsin can be inactivated by the oxidants in cigarette smoke which convert methionine to methionine sulphoxide. Researchers are experimenting with a mutant protein, a human-made family member, in which the methionine is replaced with an amino acid resistant to oxidation. It could possibly be administered as an aerosol and be useful for smokers and for those with defective α_1-antitrypsin.

Metaphors & GTPase

Before continuing this tour of gene families, I want to mention something that struck me as I was investigating this topic: It seems to bring out the poet in the molecular biologist. A few years ago I wrote an article on the use of metaphors in biology (See *ABT*, February 1988). That exercise made me more sensitive to metaphors in science, but rarely have I run across a topic so laden with them. The very idea of a gene family is metaphorical. Family is the quintessential form of relationship; it is itself genetic. It implies both close ties and similarities in traits. In describing the GTPase superfamily, Bourne, Sanders and McCormick (1990) extend this metaphor:

> The time is ripe for reviewing these GTPase proteins. It may never again be possible to capture them in a family portrait, because the family multiplies so rapidly and individual family members refuse to sit still. We now know several family members well enough however, to recognize their cousins anywhere.

This is a beautiful quote that borders on the poetic. The article does indeed provide a detailed family portrait, identifying five branches to the family tree (more metaphors). In all these proteins, GTP binds and is hydrolyzed to GDP and inorganic phosphate. The differences are in how the hydrolysis is coupled to other functions, in how the energy of hydrolysis is used. One branch of the family is composed of elongation and initiation factors for ribosomal protein synthesis. Another group is essential to the control of cell proliferation and differentiation, a third to guiding the movement of vesicles within cells and a fourth to translocation of newly synthesized proteins into the endoplasmic reticulum. The final branch includes proteins involved in transmembrane signaling. In another metaphorical moment, the authors note that:

> Compared with other protein signalling machines, the G proteins stand out as virtuosos at sorting and amplifying transmembrane signals. Each G protein determines the flow of information from a specific subset of receptors to a similarly specific and usually smaller subset of effector molecules.

Tinkering

Another powerful metaphor concerning the development of gene families is used by François Jacob (1982). He argues that engineering metaphors do not adequately describe natural selection. Engineers start from scratch, they build from plans and they produce the best results possible with the technology of the time. Jacob sees evolution as quite different from this. A new organism is not created *de novo*; there is no preconceived plan and, he notes, no perfection in nature. He argues that evolution is more like tinkering than engineering. Someone who tinkers:

> manages with odds and ends. Often without even knowing what he is going to produce, he uses whatever he finds around him, old cardboards, pieces of string, fragments of wood or metal, to make some kind of workable object. . . . In many instances, and without any well-defined long-term project, the tinkerer picks up an object which happens to be in his stock and gives it an unexpected function.

Jacob sees similar processes going on in evolution: "Darwin showed how new structures are elaborated out of pre-existing components, which initially were in charge of achieving a given task but became progressively adopted to different functions."

The title essay in Stephen Jay Gould's (1980) book *The Panda's Thumb* describes just such a case. The panda's thumb is really not a thumb like ours at all. The panda has five digits which form the paw and its "thumb" is constructed from the radial sesamoid bone which is usually part of the wrist. In the panda, this bone is enlarged, elongated and equipped with a muscle that is usually attached to the first digit, the true thumb. Gould writes than an opposable thumb for stripping the leaves from the panda's only food, bamboo, would have been:

> an engineer's best solution, [but it] is barred by history. The panda's true thumb is committed to another role, too specialized for a different function to become an opposable, manipulating digit. So the panda must use parts on hand and settle for an enlarged wrist bone and a somewhat clumsy, but quite workable, solution. . . . It is a contraption, not a lovely contrivance. But it does its job and excites our imagination all the more because it builds on such improbable foundations.

Molecular Tinkering

But what do pandas' thumbs have to do with gene families? It turns out that the same kind of evolutionary tinkering that went into the development of the panda's thumb also goes on at the molecular level. In fact Jacob argues that it is at the molecular level that the tinkering aspect of evolution is most apparent. To appreciate what he means we need to go into a little more detail on gene structure and on how families form. As mentioned earlier, most of the genes in eukaryotes are made up of two types of sequences: exons or sequences that code for amino acids and introns or intervening sequences that do not. The latter are spliced out of the RNA before it is translated into proteins. When these introns were first discovered in the late 1970s, they took molecular biologists by surprise. It had been assumed that eukaryotic genes were like prokaryotic genes with the entire DNA sequence coding for protein. (But just this week a report was published that there are prokaryotic introns as well [Barinaga 1990].)

As introns were discovered in one gene after another and were soon found to be almost universal in eukaryotes, the question arose as to their function. Something that common had to be adaptive, providing some advantage. Some researchers speculated that introns carried information for the proper functioning of the mRNA and that translation would not occur properly without them. In some cases, this has turned out to be true. Others saw the introns as molecular garbage arising from mechanisms to duplicate sequences and of little value to the cell. With our present knowledge this hypothesis has proved difficult to prove or disprove. Walter Gilbert proposed a third explanation for the intron's ubiquity: Intervening sequences, by breaking genes into pieces, make tinkering easier. The pieces, the exons, can be moved around and put together in new combinations to form new proteins with new functions.

Stanley and Luzio (1988) describe this process metaphorically as "evolutionary cut-and-paste," creating proteins with a "mosaic character." They are referring specifically to a family of cytolytic proteins produced by white blood cells. One of these proteins is perforin, which is made by killer T cells and can form pore-like lesions in a target cell. Other members of the family are components of the complement system, specifically C7, C8 and C9. All these proteins have similar amino acid sequences in the portion of the molecule that interacts directly with the cell membrane. It is this similarity that puts these proteins in the same family. But there is really a family within a family here, because the complement proteins share other sequences that relate them to still other families. (The family relationships in genetics can be as complicated as those in soap operas.) One sequence is similar to a portion of the low density lipoprotein receptor and another to thrombospondin, a clotting protein. Each of these sequences is called a domain—still another metaphor. Often a domain corresponds to one exon, one coding region in a gene, which lends support to Gilbert's hypothesis about new proteins being created by exon shuffling.

The Exon Universe

Recently, Gilbert and his colleagues at Harvard University have attempted to answer the question: "How big is the universe of exons?" (Dorit, Schoenbach & Gilbert 1990). They reasoned that, "If genes have been assembled from exon subunits, the frequency with which exons are reused leads to an estimate of the size of the underlying exon universe." By looking at known protein sequences and using a rather sophisticated and controversial computer program, Gilbert estimates that, "Only 1000 to 7000 exons were needed to construct all proteins." While some experts debate Gilbert's approach and his specific estimates, all agree that gene families present a beautiful example of the underlying unity found in the diversity of life.

Just as in a Bach fugue, where the same theme, the same motif, is used over and over again in different settings, so nature uses motifs in a similar way.

There seems to be a constant process of mix-and-match in nature, and while much of this genetic shuffling produces nothing useful, a few combinations prove highly functional. Natalie Angier (1990) describes nature as "rather lazy, preferring to build new proteins from pre-existing parts over creating a possibly more effective working unit from scratch." Again, nature tinkers rather than engineers.

More Metaphors

Nature may also be described as an artist using the same motifs in a variety of ways. In fact, motif is another word used metaphorically in discussing gene families. A motif is a recurring element, a central theme in a work of art or music. In genetics, it refers to a sequence that recurs in different proteins. Its meaning is similar to that of domain, but motif often implies a shorter sequence. For example, there are odorant-binding proteins secreted from the nasal glands which concentrate and deliver odorants to scent receptors. Now similar proteins have been found secreted from von Ebner's glands on the tongue; these proteins may serve the analogous role of delivering flavor molecules to taste receptors (Schmale, Holtgreve-Grez & Christiansen 1990). Not only are these two groups of proteins similar to each other, they also belong to a superfamily of hydrophobic molecule transporters which carry retinoic acid, the steroids and other ligands.

Besides other similarities, all members of this superfamily share highly conserved "motifs" just four amino acids long, which are important in membrane interactions. Just as in a Bach fugue, where the same theme, the same motif, is used over and over again in different settings, so nature uses motifs in a similar way. Perhaps part of our attraction to Bach is because his reiterative methods were also used in the creation of our own genome. This may seem rather mystical, but in his *Gödel, Escher, Bach: An Eternal Golden Band*, Douglas Hofstadter (1979) discusses just such similarities, not only between a Bach fugue and a gene sequence, but in the repetitive drawings of M.C. Escher, in the reiterative methods of processing information in the brain and in the idea of self-reference in Gödel's Theorem in mathematics. The fact that such similar patterns arise in such different contexts argues for the fundamental nature of the concept of variations on a theme, of unity in variety.

A number of the metaphors used to describe the evolution of gene families are taken from other aspects of evolution. For example, some sequences, like some macroscopic traits, are highly conserved over a wide range of evolutionarily diverse organisms: Not only are biochemical pathways very similar in very different organisms, but the amino acid sequences in the pathways' enzymes are very similar. That domains and motifs are conserved implies the evolutionary concept of constraint; little change is possible because most change would create a less effective molecule. There is little room for diversification; it would be nonadaptive. Still another evolutionary concept that is relevant here is that of convergence. In some cases, it appears that sequence similarities between two proteins exist not because the proteins share a common ancestor, but because they have experienced similar adaptive pressures. For example, lipases and serine proteases do not belong to the same gene family, they aren't even cousins, but they all have a similar amino acid configuration around the active site: aspartic acid . . histidine . . serine. David Blow (1990) argues that this similarity arose by coincidence and was conserved because of its enzymatic activity.

Another example of convergence is particularly interesting; it involves both anatomic and genetic convergence. The squid's eye has an optic nerve, lens, pupil, cornea, iris and retinal receptor cells as does a vertebrate eye and the elements are organized similarly in both. Yet these eyes evolved independently of each other, providing one of the most spectacular examples of convergent evolution. The lens protein in the squid eye, S-crystallin, also has very similar properties to the crystallins in vertebrates. But again, this is the result of convergence, not descent from a common ancestor. S-crystallin does belong to a family of proteins, but a family that, rather than including other crystallins, includes the mammalian enzyme glutathione S-transferase. And other crystallins also belong to enzyme families. ξ-crystallin of the guinea pig lens is related to alcohol dehydrogenase and frog ρ-crystallin to aldose reductase. Russell Doolittle (1988) argues that the properties of crystallins are such that many types of proteins can fit the bill. He sees this as an example of—to use another metaphor—molecular opportunism:

> Crystallins are . . . water-soluble proteins that can pack together efficiently to form very large molecular aggregates, and it could be that the only requirement for a structural lens protein is for a globular protein that can pack well and form transparent aggregates.

Just the Beginning

With all these comparisons of DNA and amino acid sequences, with all this work on protein genealogies, we have to keep reminding ourselves that the construction of protein family trees is still in its infancy. There are so many proteins and genes that have yet to be identified and sequences to be discovered that the hunt for family members has really only just begun. For some families like the immunoglobulin family, extensive genealogies have been worked out. In other cases, new families are still being identified. A March 1990 article on the enzymes that attach amino acids to transfer-RNAs, the aminoacyl-t-RNA synthetases, describes them as all belonging to one family (Moras 1990). Six months later, when the structure of another synthetase was worked out, it was found to be very different from the previously studied molecules (Cusack et al. 1990). There doesn't seem to be any evolutionary relationship between the two types, so a new family of synthetases has been born. Researchers have also recently sequenced the gene for a kinase that phosphorylates neurotransmitter receptors only when they are coupled to neurotransmitter. This enzyme might play a role in regulating receptor activity, and it is the first sequenced member of what may become a large family of receptor kinases (Benovic et al. 1989).

The discovery of new families and new family members obviously adds to the interest of genetic genealogy. Also appealing is the idea that this study can allow us a glimpse into the distant molecular past. We can see how proteins and genes changed over time. Those who study the origin of words are drawn by similar attractions: finding where our words and the roots of language originated. Lewis Thomas (1990) is a noted physician and immunologist who has just published a book, *Et Cetera, Et Cetera: Notes of a Word-Watcher*, as a companion to his *Lives of a Cell: Notes of a Biology Watcher* (1974). In that earlier volume, he already displayed some of his love of words in an essay on the origin of the word "gene." In describing the process of word evolution, he writes that, "The words themselves must show the internal marks of long use; they must contain their own inner conversation." The same can very well be said of our genes, and we have only begun our study of genetic philology—to coin a new metaphor.

References

Angier, N. (1990, December 11). Nature may fashion all cells' proteins from a few primordial parts. *The New York Times*, pp. C1,C9.

Barinaga, M. (1990). Introns pop up in new places—what does it mean? *Science, 250*, 1512.

Benian, G., Kiff, J., Neckelmann, N., Moerman, D. & Waterson, R. (1989). Sequence of an unusually large protein implicated in regulation of myosin activity in *C. elegans*. *Nature, 342*, 45–50.

Benovic, J., DeBlasi, A., Stone, W., Caron, M. & Lefkowitz, R. (1989). β-Adrenergic receptor kinase: Primary structure delineates a multigene family. *Science, 246*, 235–240.

Blow, D. (1990). More of the catalytic triad. *Nature, 348*, 694–695.

Bourne, H., Sanders, D. & McCormack, F. (1990). The GTPase superfamily: A conserved switch for diverse cell functions. *Nature, 348*, 125–132.

Carrell, R. (1984). Therapy by instant evolution. *Nature, 312*, 14.

Cusack, S., Berthet-Colominas, C., Härtlein, M., Nasser, N. & Leberman, R. (1990). A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 Å. *Nature, 347*, 249–255.

Doolittle, R. (1988). More molecular opportunism. *Nature, 336*, 18.

Dorit, R. Schoenbach, L. & Gilbert, W. (1990). How big is the exon universe? *Science, 250*, 1377–1382.

Evans, R. (1988). The steroid and thyroid hormone receptor family. *Science, 240*, 889–895.

Gould, S.J. (1980). *The panda's thumb*. New York: Norton.

Hofstadter, D. (1979). *Gödel, Esher, Bach: An Eternal Golden Band*. New York: Basic.

Jacob, F. (1982). *The possible and the actual*. New York: Pantheon.

Maeda, N. & Smithies, O. (1986). The evolution of multigene families: Human haptoglobin genes. *Annual Review of Genetics, 20*, 81–108.

Moras, D. (1990). Synthetases gain recognition. *Nature, 344*, 195–196.

Nebert, D. (1990). Growth signal pathways. *Nature, 347*, 709–710.

Schmale, H., Holtgreve-Grez, H. & Christiansen H. (1990). Possible role for salivary gland protein in taste reception indicated by homology to lipophilic-ligand carrier proteins. *Nature, 343*, 366–369.

Stanley, K. & Luzio, P. (1988). A family of killer proteins. *Nature, 334*, 475–476.

Thomas, L. (1974). *The lives of a cell: Notes of a biology watcher*. New York: Viking.

Thomas, L. (1990). *Et cetera, et cetera: Notes of a word-watcher*. Boston: Little, Brown.