ARTHUR M. LESK

The Unreasonable Effectiveness of Mathematics in Molecular Biology*

My title is an emulation of that of the well-known paper by E.P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences [1]." Of course the irony cuts in opposite ways in physics and molecular biology. In physics, mathematics is obviously effective— many of the giants on whose shoulders physicists stand are mathematicians—and the surprise is Wigner's suggestion that this is unreasonable. In molecular biology, the proper role of mathematics is not obvious, and there is fear, far more credible than for physics, that it may be unreasonable to expect mathematics to be effective. Of course, many common *tools* of computational molecular biology—for instance, searching in databases for sequences similar to a probe sequence—are certainly based on mathematics and computer science. But whether our ultimate understanding of living processes will be expressed in the language of mathematics—in the way, for example, that concepts of symmetry underlie the statement of laws of physics—or in the traditional descriptive "anecdotal" language of biology, is still moot.

Why might it be reasonable to doubt the effectiveness of mathematics in biology? Observed properties of living systems are determined by a combination of

- The laws of physics and chemistry
- The mechanism of evolution
- Historical accident

It is difficult to sort out their effects, and a creative tension among them pervades our investigations. Many of the laws of physics describe the natural world—including living systems—by specifying relations between initial and fi-

*Based on a talk delivered at the final symposium of the program, "Biomolecular Function and Evolution in the Context of the Genome Project," at The Isaac Newton Institute for the Mathematical Sciences, Cambridge, U.K., 20 Dec. 1998.

nal conditions. In biology the complexity of the set of possible initial conditions creates difficulties. The large role of historical accident hinders and humbles us: Even if fundamental laws of physics and chemistry have simple consequences that would provide detailed descriptions of life processes, we may not be able to discover them, because our observables are more complex, resist simplifying idealizations, and show features dominated by a choice of initial conditions from a very large and diverse set of possibilities. Apples, in biology, do much more than fall on people's heads.

The Subject Matter of Computational Molecular Biology

The objects of our study at least have a form to which we can attempt to apply mathematics. These include:

- DNA sequences of genes
- amino acid sequences of proteins
- protein structures
- protein functions

Readers will have heard of the *genome* projects, which are the determinations of the complete sequences of the DNA in organisms: the set of blueprints. The DNA sequences in genomes contain all the information required by the organism to be born, to develop and grow, and to die. With the completion of the sequencing of the yeast genome in 1996, we know as much about a yeast cell as a yeast cell does. This statement is not as arrogant as it sounds. We *do* possess all the information. Admittedly, we may not be able to interpret it as effectively as a yeast cell can, but we do have the complete set of blueprints. But blueprints give only a *static* description of structure and potential activity; it remains to extend our observations to the integration of protein expression and function, in time and space within an organism. Collection of these data is known as the "proteome project," and it is gathering momentum for a role in the post-genomic era.

The rate of measurement of gene sequences is extremely large, and increasing. In 1998 the complete sequence of the DNA from a worm, *Caenorhabditis elegans*, was completed (9.7×10^7 bases). It is likely that 1999 and 2000 will see, respectively, the completion of the sequencing of the DNA from the fruit fly (1.8×10^8 bases) and the human genome (3.4×10^9 bases), as well as numerous other organisms both large and small. Louis XV could say, "Après moi, le déluge." Noah, in contrast, could not; nor can we.

The sequences and structures that we study are interrelated in important ways. On the molecular level, the DNA sequences of genes encipher the amino acid sequences of proteins. Then the amino acid sequences of proteins determine the three-dimensional structures of proteins. Protein structure then determines protein function (Figure 1). A precise three-dimensional structure is necessary for protein function because the required interactions depend on bringing together different parts of molecules in exact spatial relationships. Finally, the feedback from protein function back to gene sequence—by evolution through natural selection—closes the loop.

In our computers DNA sequences are character strings: one-dimensional objects. Genes, substrings of genome sequences, are translated into amino acid sequences of proteins, by a nearly-universal cipher. Amino acid sequences of proteins are also represented by one-dimensional character strings. Next, proteins fold spontaneously to unique "native" three-dimensional structures. (The evidence for this is that they can be denatured—the three-dimensional structure destroyed—by heating, and when cooled they resume their original form, like shape-memory alloys.) The spontaneous folding of proteins is the point at which Nature makes the leap from the one-dimensional sequences of genes to the three-dimensional world we inhabit.

The Goals of Computational Molecular Biology

What are our goals in dealing with this material? First, simply to describe the similarities and differences among sequences and among structures, and to classify them. What are the topologies of sequence space, structure space, and the space of protein function? What are the mappings among these spaces? We wish to be able to describe and predict the interrelationships among protein sequence, structure, and function, using evolution as the organizing principle.

How do we go about this work? Sydney Brenner once lamented: "The problem with biology is that it has no harmonic oscillator." By this he meant that in biology, unlike physics, there is no escape from complexity, even *via* idealization. In physics, the harmonic oscillator is a simple problem, solvable exactly by many methods; it applies exactly to some phenomena, and is a useful approximation to others. In physics it is the traditional testbed for new methods. In fact, *computational* molecular biology has two "harmonic oscillators" in Brenner's sense: Sequence alignment, and Structure superposition. These operations, which can be carried out exactly and efficiently, are basic to many analyses of sequence–structure relationships in molecular biology. Now, it will come as no surprise that the real world is often anharmonic. Nevertheless, many valuable tools have been developed from these simple cases.

However, tools provide answers but not questions. Research in this field continues to depend on the interac-

The leading mathematician I.M. Gelfand, who is also a leading physiologist, hotly denies being a mathematical biologist. To Wigner's Principle—the unreasonable effectiveness of mathematics in the physical sciences—he counterposes

the unreasonable ineffectiveness of mathematics in the biological sciences

which some of his colleagues advocate calling the Wigner-Gelfand Principle.

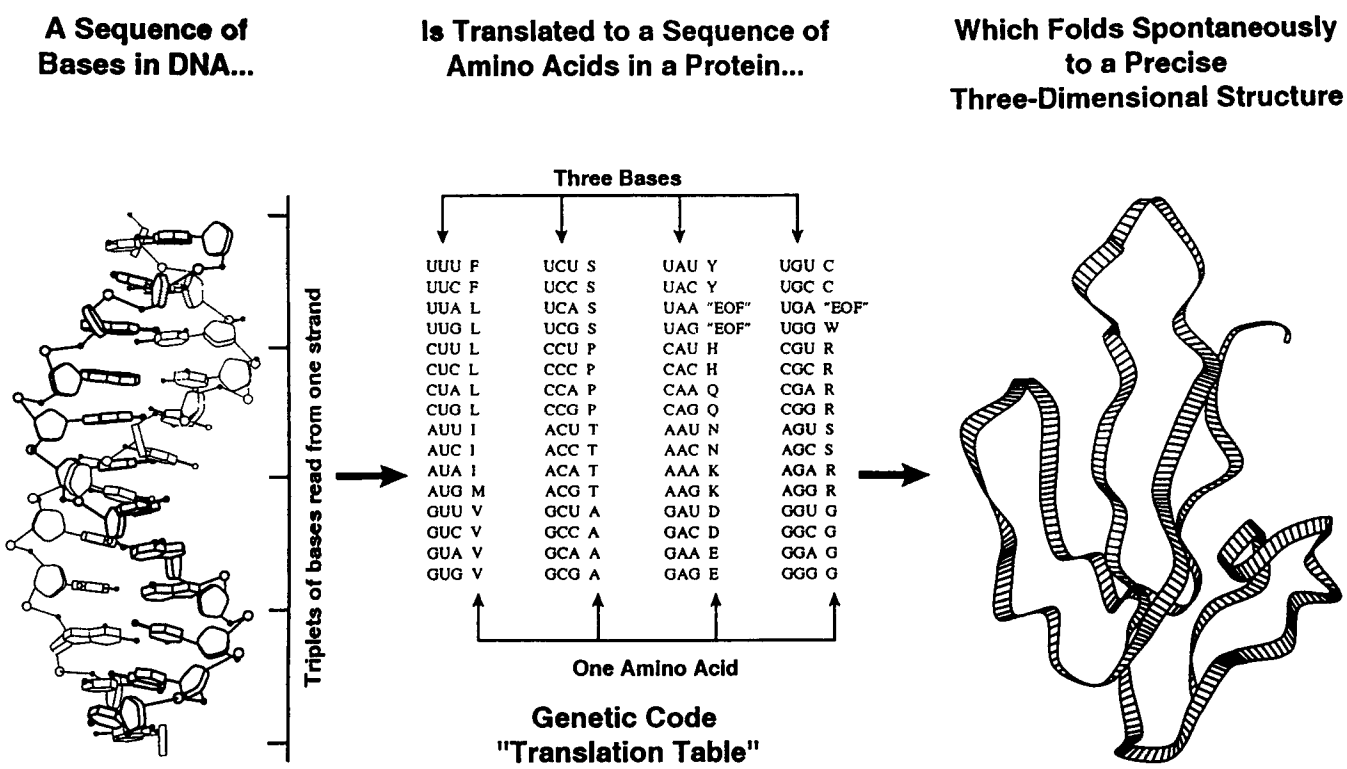

Figure 1. Information flow during "readout" from a gene. ***Genes, the basic blueprints of organisms, are contained in the structure of DNA (left).*** **At the left of the figure is the double helix of DNA, containing two intertwining strands—one drawn in narrow lines, the other in bold. The "staircase effect" is created by a set of chemical subunits called the "bases". At each position on either strand there are four possible bases: A, T, G, and C. The sharp-eyed reader can see that the bases—the "risers" of the helical staircase—come in different forms. Each base interacts with another base at the same level, on the other strand, and these interactions demand strict complementarity: presence of an A on one strand requires a T opposite it on the other, a G on one strand requires a C on the other, and vice versa: a T on one strand is complementary to an A and a C to a G. In this way, each strand contains enough information to direct the synthesis of its partner. Logically, the way to replicate DNA is to take the strands apart and synthesize the complement of each of the separated strands.** ***The sequences of bases in genes encipher the amino acid sequences of proteins, by a direct translation table known as "the genetic code" (center).*** **The genetic message is written in the four-character A, T, G, C alphabet. Proteins are also polymers containing a sequence of chemical residues such that each position contains one of *twenty* amino acids, with mnemonics A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. To specify a total 20 amino acids, each requires more than two bases; in fact the DNA sequence is read three bases at a time, with a redundancy that is quite important for evolution. (Three triplets of bases are reserved for "End-of-file" terminator signals.)** ***Proteins fold spontaneously to native, active three-dimensional structures (right).*** **This is the point at which Nature makes the leap from the one-dimensional sequences of genes to the three-dimensional world we inhabit. The example shown is a toxin from a sea snake, one of the many protein structures determined by X-ray crystallography. Each gene has a sequence of bases that dictates first the amino acid sequence of a protein, thereby its three-dimensional structure, and thereby its function.**

tion of the human scientist with the data, using mathematical and computational methods in support.

Sequences and Sequence Alignment

Gene and protein sequences take the form of character strings. For gene sequences the characters are chosen from a set of four, {A, T, G, C}, symbolizing the nucleotides Adenine, Thymine, Guanine, and Cytosine. For protein sequences the characters are chosen from a set of twenty, symbolizing twenty canonical amino acids.

Alignment

The alignment of two character strings is the determination of a meaningful correspondence between their elements.

Given two character strings:

g c t g a a c and c t a t a a t c

then two possible alignments are:

```
g c t g a - a - - c     and     g c t g - a a - c
- c t - a t a a t c             - c t a t a a t c
```

How can we decide which of these two, or of many other possibilities, is the best alignment? Can we devise a metric for character strings and define distances between them? Measures of dissimilarity between character strings include:

(1) The *Hamming* distance, defined between two strings of equal length, is the number of positions with mismatching characters.

(2) The *Levenshtein* distance, between two strings of not necessarily equal length, is the minimal number of "edit operations" required to change one string into the other, where an edit operation is a deletion, insertion, or alteration of a single character in either sequence. A given sequence of edit operations induces a unique alignment, but not vice versa.

In molecular biology, we know that insertions and deletions have occurred in gene and protein sequences. Therefore the Hamming distance is not general enough. Moreover, there is evidence that some changes are more likely to have occurred than others. Even the Levenshtein distance must therefore be generalized, to include differential weighting of different edit operations, based on our underlying evolutionary model. For instance, mutations are likely to be conservative: the replacement of one amino acid in a protein by another with similar size or physicochemical properties is more likely than its replacement by another amino acid with more dissimilar properties. To reflect this, instead of a discrete counting of edit operations we assign a "cost" ($\in \mathbb{R}$) to each change in the sequence. Also, there is evidence that the cost of a gap is not proportional to its length as in the Levenshtein model; however, the proper choice of gap weights as a function of length is a matter of considerable delicacy. Many schemes apply a linear function with one parameter α for gap initiation and another, smaller, parameter β for gap extension, to give a gap cost of the form $\alpha + \beta \times (gaplength - 1)$. Algorithms exist to determine best alignments by minimizing the sum of the costs of the edit operations that transform one string into the other.

A formal statement of the problem of optimal sequence alignment is as follows: We are given two character strings: $A = a_1a_2\cdots a_n$ and $B = b_1b_2\cdots b_m$, with each a_i and b_j a member of an alphabet set $\mathcal{A}$. Let $\mathcal{A}^+ = \mathcal{A} \cup \{\phi\}$. A sequence of edit operations is a set of ordered pairs (x, y), with $x, y \in \mathcal{A}^+$. Individual edit operations include:

Substitution of b_j for a_i—represented (a_i, b_j).
Deletion of a_i from sequence A—represented (a_i, ϕ).
Deletion of b_j from sequence B—represented as (ϕ, b_j).

A *cost function* d is defined on edit operations:

$d(a_i, b_j)$ = cost of a mutation
$d(a_i, \phi)$ or $d(\phi, b_j)$ = cost of a deletion or insertion.

The minimum weighted distance between sequences A and B is

$$D(A, B) = \min_{A \to B} \Sigma d(x, y),$$

where $x, y \in \mathcal{A}^+$ and the minimum is taken over all sequences of edit operations that convert A to B. If $d(x, y)$ is a metric on $\mathcal{A}^+$, $D(A, B)$ is a metric on strings of characters from $\mathcal{A}^+$. (This statement of the problem assumes length-independent gap costs; more realistic gap-weighting schemes are generalizations.)

The problem is to find $D(A, B)$ and one or more of the alignments that correspond to it. An algorithm that solves this problem in $\mathbb{O}(mn)$ time has been known for a long time, and has been applied to many problems including text editing, speech recognition, and analysis of birdsongs [2]. It entered molecular biology in a seminal paper by Needleman and Wunsch [3].

Several features of this algorithm are noteworthy.

- It produces a *global* optimum: recall that this is the first of computational molecular biology's two "harmonic oscillators." We have a method guaranteed not to get trapped in local minima.
- That was the good news. The bad news is that interpretation of the results is not so straightforward. Although a sequence of edit operations derived from an optimal alignment *may* correspond to an actual evolutionary pathway, it is impossible to prove that it does. The larger the edit distance, the larger the number of reasonable evolutionary pathways. Not only may the optimal alignment be nonunique, but there may be many suboptimal alignments that score quite close to the optimal. For example, Fitch and Smith analysed the chicken genes for α and β haemoglobin [4]. They found 17 optimal alignments, one of which agreed with the alignment based on the known haemoglobin structures, and over a thousand alignments scoring within 5% of optimum.

Problems with Pairwise Sequence Alignment

It is observed that as proteins evolve, the amino acid sequence diverges more quickly than the structure. In many cases we can detect an evolutionary relationship between two protein structures even though there is no detectable similarity between the gene sequences or the amino acid sequences. What is happening is this: The genes are exploring the space of DNA sequences, but natural selection is acting as a brake on changes in structure, in order to conserve function. The redundancy in the genetic code—the facts that several different triplets of bases encipher the same amino acid, and that many single-base changes interconvert amino acids with similar physico-chemical properties—moderates the structural consequences of sequence changes.

Even if similarity *is* detectable at the sequence level, for distantly-related proteins the optimal pairwise sequence alignment often gives the wrong answer, relative to comparison of the structures, which is the court of last resort.

However, if many related sequences are available, multiple sequence alignment gives more significant and accurate results than pairwise sequence alignment. Why do multiple alignments enhance sequence information? It is the appearance of patterns of conservation. The extent and nature of the variation at individual positions is an important guide to the structural or functional role of different regions of the sequence (Figure 2). For instance, residues conserved over an entire family of proteins are usually involved in function, or at least are usually essential for the structure. Conversely, regions in which insertions and deletions are very common usually correspond to peripheral regions.

```
TYLWEFLLKLLQDR.EYCPRFIKWTNREKGVFKLV..DSKAVSRLWGMHKN.KPD
VQLWQFLLEILTD..CEHTDVIEWVG.TEGEFKLT..DPDRVARLWGEKKN.KPA
IQLWQFLLELLTD..KDARDCISWVG.DEGEFKLN..QPELVAQKWGQRKN.KPT
IQLWQFLLELLSD..SSNSSCITWEG.TNGEFKMT..DPDEVARRWGERKS.KPN
IQLWQFLLELLTD..KSCQSFISWTG.DGWEFKLS..DPDEVARRWGKRKN.KPK
IQLWQFLLELLQD..GARSSCIRWTG.NSREFQLC..DPKEVARLWGERKR.KPG
IQLWHFILELLQK..EEFRHVIAWQQGEYGEFVIK..DPDEVARLWGRRKC.KPQ
VTLWQFLLQLLRE..QGNGHIISWTSRDGGEFKLV..DAEEVARLWGLRKN.KTN
ITLWQFLLHLLLD..QKHEHLICWTS.NDGEFKLL..KAEEVAKLWGLRKN.KTN
LQLWQFLVALLDD..PTNAHFIAWTG.RGMEFKLI..EPEEVARLWGIQKN.RPA
IHLWQFLKELLASP.QVNGTAIRWIDRSKGIFKIE..DSVRVAKLWGRRKN.RPA
RLLWDFLQQLLNDRNQKYSDLIAWKCRDTGVFKIV..DPAGLAKLWGIQKN.HLS
RLLWDYVYQLLSD..SRYENFIRWEDKESKIFRIV..DPNGLARLWGNHKN.RTN
IRLYQFLLDLLRS..GDMKDSIWWVDKDKGTFQFSSKHKEALAHRWGIQKGNRKK
LRLYQFLLGLLTR..GDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKR

  L  fl  lL          i W       F          a  WG  K
```

Figure 2. Multiple alignment of partial sequences from a family of proteins called ETS domains. Each line corresponds to the amino acid sequence from one protein, specified as a sequence of letters each specifying one amino acid. Looking down any column shows the amino acids that appear at that position in each of the proteins in the family. In this way patterns of preference are made visible. For instance, the third position is a leucine, L, in every sequence—this implies that some structural or functional constraint has dissuaded evolution from changing this position. Letters underneath the table indicate positions of invariant residues (upper case) or invariant with one exception (lower case). Note the uneven distribution of variability in the different columns. The periodicity of conserved residues (3, 4 or 8) suggests that the proteins contain helices, which is true. Other patterns are hidden more deeply, and require computational analysis to identify them. Such patterns might include correlations between the distributions of amino acids at different positions: For instance, in the fourth column from the left the amino acid tyrosine, Y, appears only in the last two sequences; the others contain tryptophan, W. There is approximate correlation in the pattern of variability between this column and the fourth and fifth columns from the right. It is widely believed (or at least hoped) that *correlations* of patterns of variability at different positions of such a table of sequences should give some clues about sites that interact in three dimensions. Unfortunately the signal is quite weak.

(One sequence plays coy about its structure. A pair of aligned sequences whisper about their structure. Three or more sequences shout about their structure out loud.)

If sequences give only an indirect glimpse of structures, why not deal with structures alone? The reason is that the amount of sequence data known far exceeds the amount of structural data. For about 20 organisms the entire genome is sequenced, giving us the complete sequences of all the genes. Only for a small minority of these genes do we have the three-dimensional structures of the corresponding proteins.

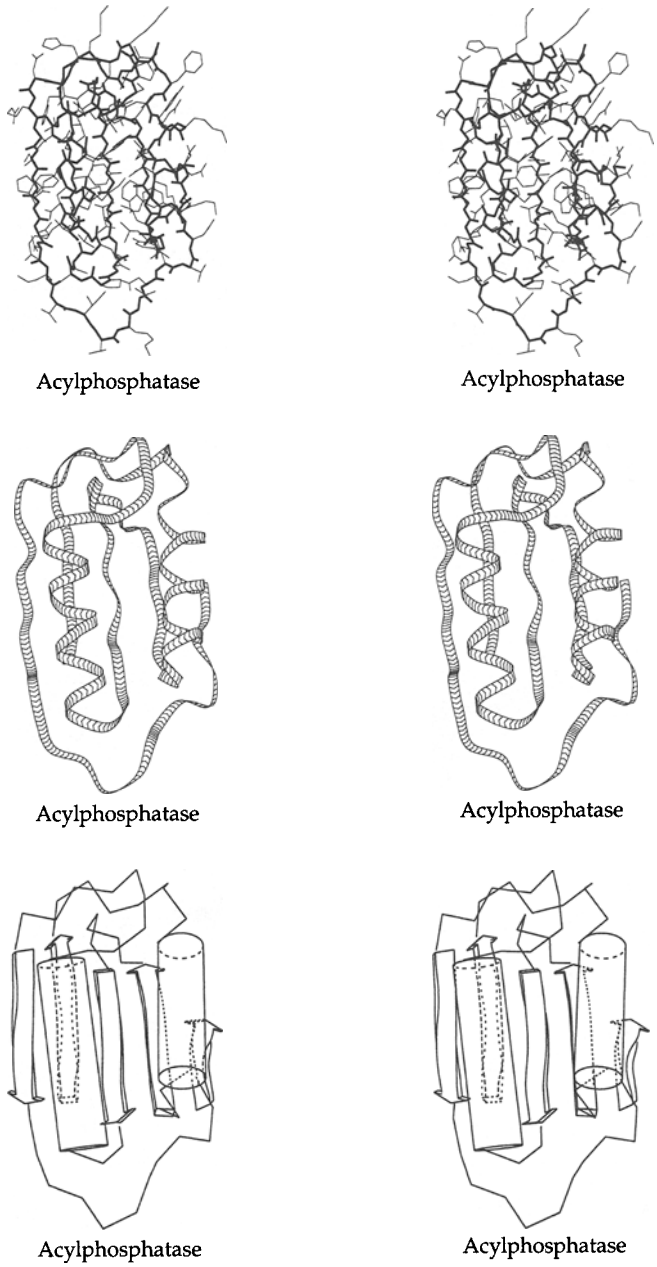

Figure 3. Proteins are sufficiently complex structures that it has been necessary to develop specialized tools to present them. This figure shows a relatively small protein, acylphosphatase, at three different degrees of simplification. Top: complete skeletal model; mainchain bolder than sidechains. Center: the course of the chain is represented by a smooth interpolated curve, the chevrons indicating the direction of the chain. Bottom: schematic diagram, in which cylinders represent helices and arrows represent strands of sheet. The solid objects in the picture are represented as "translucent" by altering lines that pass behind them to broken lines. To superpose adjacent representations, rotate the page by 90° and view in stereo (not for too long!).

Analysis of Protein Structures

The first problem in analyzing the structures of molecules as complex as proteins is one of presentation. Computer-graphic techniques have been developed to draw simplified representations of proteins. Figure 3 illustrates, for a small protein molecule, the difficulty in interpreting a fully detailed, literal representation, and the kind of simplified pictures that programs produce to give us visual access to the material. An active cottage industry has produced many different representations; that is, many people have proposed different simplified representations, and these become subsumed into general graphics packages. A skilled

molecular illustrator will combine them to show different aspects of a structure in finely-tuned degrees of detail. Such pictures, rendered in full color and with fancy (but unrealistic, given the size of the molecules relative to the wavelength of visible light) shadowing effects, adorn journals, posters, and even T-shirts and mugs.

We now know the structures of 10000 proteins, and see in them a vast variety of spatial patterns. To Rutherford's comment, "All science is either physics or stamp collecting," I reply that the study of protein structure combines the best of both fields! We have the spectacular variety, but also have faith in the existence of underlying unifying principles.

Every protein consists of a linear (that is, unbranched) repetitive polymer mainchain with different amino acid sidechains hung on it at regular intervals. A protein is analogous to a string of Christmas tree lights, with the wire corresponding to the repetitive mainchain, and the sequence of colors of the lights to the individuality of the sequence of sidechains.

The mainchain describes a space curve which is stabilized by favourable interactions between the sidechains that are brought into contact. Such a space curve is most easily seen in the center frame of Figure 3. Two regions at the front of the picture have the form of *helices*—they look like a classic barber pole—with their axes almost vertical. This is one of the two standard structures that local regions of the chain adopt. The other standard structure is the almost extended strand of *sheet:* the protein in Figure 3 contains four strands of sheet, approximately vertical in orientation. These strands interact laterally to stabilize their assembly. In the bottom frame of Figure 3, the helices and strands are represented as "icons": helices as cylinders and strands of sheet as large arrows. The top frame of Figure 3 shows the most detailed representation of the structure, including mainchain and sidechains; the contrast demonstrates the importance of simplification in producing a visually intelligible picture of even a small protein.

The initial stage of "parsing" a new structure involves identifying the regions of helix and sheet. This is the information required to convert the representation in Figure 3, center frame, to that of the bottom frame. The most common type of helix in protein structures contains 3.6 residues per turn. Features in the *sequence* that show this periodicity suggests helical regions.

Superposition of Structures

As in the case of sequences, a fundamental question in analyzing structures is to devise and compute a measure of similarity. Suppose that we have coordinate sets representing two structures:

$$p_i = (x_i, y_i, z_i), \qquad i = 1, \ldots N$$

and

$$q_j = (x'_j, y'_j, z'_j), \qquad j = 1, \ldots M.$$

Just as in the case of sequences, the question of alignment arises. Consider the contrast between three related problems that arise in computational chemistry:

(1) *Measure the similarity of two sets of atoms with known correspondences:*

$$p_i \longleftrightarrow q_i \qquad i = 1, \ldots N.$$

(The analog for sequences is the Hamming distance.) This problem can be solved exactly and efficiently—it is the second "harmonic oscillator" of computational molecular biology.

(2) *Measure the similarity of two sets of atoms with unknown correspondences, but for which the molecular structure—specifically the linear order of the residues—restricts the correspondence.* In the case of proteins the alignment must retain the order along the chain:

$$p_{i(k)} \longleftrightarrow q_{j(k)}, \qquad k = 1, \ldots, K \leq N, M,$$

with the constraint that $k_1 > k_2 \Rightarrow i(k_1) > i(k_2)$ and $j(k_1) > j(k_2)$. This can be thought of as corresponding to the Levenshtein distance between character strings, or to sequence alignment with gaps.

(3) *Measure the similarity between two sets of atoms with unknown correspondence, with no restrictions on the correspondence:*

$$p_{i(k)} \longleftrightarrow q_{j(k)}, \qquad k = 1, \ldots, K \leq N, M$$

This problem arises in the following important case: Suppose two (or more) molecules have similar biological effects, such as a common pharmacological activity. It is often the case that the structures share a common constellation of a relatively small subset of their atoms that is responsible for the biological activity, called a *pharmacophore.* To identify the pharmacophore it is useful to be able to find the maximal subsets of atoms from two or more molecules that have similar structure.

Problems (2) and (3) require determination of the alignment of the points. Alignment methods based exclusively on the coordinates (that is, not on the amino acid sequences) are called *structural alignments.* In structural alignments, corresponding residues are identified because they occupy the same position relative to the structure as a whole. One must think of extracting the maximal common substructure and basing the alignment on this. (For instance, the maximal common substructure of the letters B and R is the letter P.) Residues outside the maximal common substructure are unalignable, a fact that cannot be detected by pairwise sequence alignment; this is one of its weaknesses.

The most general approach to these three problems is based on the solution of problem (1), the case of known correspondence $p_i \longleftrightarrow q_i$. Two *identical* objects can be superposed by a rigid-body translation and rotation of one of them onto the other. Two objects that are *similar* can be brought into *approximate* superposition by rotation and translation. If the objects are ordered sets of points, a measure of their similarity is the root-mean-square deviation Δ after optimal superposition:

$$\Delta^2 = \min_{\mathbf{R},\mathbf{t}} \{\sum_{i=1}^{N} \| \mathbf{R}p_i + \mathbf{t} - q_i \|\},$$

where $\mathbf{R}$ is a proper rotation matrix (det $\mathbf{R} = 1$) and $\mathbf{t}$ is a translation vector. In the optimal superposition, the mean

positions (colloquially, the "centers of gravity") of the two sets coincide. The problem of determining the correct relative orientation is known as the "Orthogonal Procrustes problem," and solutions based on standard techniques of linear algebra are available [5].

Solution of the maximal common substructure problem provides the basis of a metric for structures. It allows detection of partial and tenuous similarities, and induces a classification tree of the entire corpus of protein structures.

Approaches to maximal common substructure calculations have been based on two representations of structures: (1) in terms of lists of coordinates $p_i = (x_i, y_i, z_i)$, $i = 1, \ldots, n$, or alternatively (2) in terms of distance matrices $D_p(i,j) = |p_i - p_j|$. The main advantage of distance matrices is that they provide an origin- and orientation-independent representation of the structure. In terms of distance matrices, the maximal element of the difference distance matrix, $\max_{i,j} \{|D_p(i,j) - D_q(i,j)|\}$, provides one measure of the structural difference between two aligned point sets.

Coordinates and distance matrices are nearly equivalent representations of a point set. Calculation of the distance matrix from the coordinates is trivial. It is less obvious that the coordinates can be recovered exactly and directly from the distance matrix, but this can be done by a matrix diagonalization [6]. To be sure, the distance matrix specifies equally both the original structure and its enantiomorph (thus corresponding right and left gloves are two enantiomorphs), but this ambiguity is not a serious problem for applications to molecular biology. Position and orientation information are of course also lost.

The major difficulty in maximal common substructure calculations of types 2 and 3 is the combinatorial complexity of considering the many possible alignments. Algorithms based on distance matrices have proved more effective than those based on coordinate sets in dealing with this. Related matrix representations based on structural elements such as helices or sheets rather than on atomic coordinates provide compact representations of protein folding patterns. Extraction of maximal common submatrices reveals the largest substructures with a common folding pattern. Such representations also permit enumeration of *all* possible protein folding patterns. It is estimated empirically that all natural proteins have no more than about 1000 different folding patterns. Complete enumeration allows us to examine nature's choices, to try to distinguish between historical accident and architectural necessity.

Protein Evolution

Protein evolution is the study of how corresponding amino acid sequences and protein structures differ in related species. It is an informative type of investigation, which helps us in understanding sequence-structure relationships. For although we know that a single amino acid sequence contains all the information necessary to specify the structure of the protein, we do not yet understand how to reason from the sequence to the structure. Think of this as the "integral form" of the protein folding problem. It is an unsolved problem. In studying protein evolution we observe how *changes* in sequence are reflected in *changes* in structure; these should be easier to understand—think of this as the "differential form" of the protein folding problem:

Topic:	**Protein folding**	**Protein evolution**
Observation:	sequence → structure	change in sequence → change in structure
Form of problem:	"integral form"	"differential form"
Status of problem:	unsolved	unsolved but *should* be easier

A simple argument suggests that structure should be a nearly "continuous" function of sequence, at least for naturally evolved sequences and structures. Suppose that there were a protein such that *any* mutation (any change in the amino acid sequence) produced an unstable structure. Then, nature could not ever have achieved this structure by evolutionary processes, because it could not have had any stable precursor. It follows that natural structures must be robust. Most small changes in sequence should leave the structure intact. (This need not apply to artificially engineered protein structures.)

Indeed, natural proteins with very similar sequences have very similar structures. Before synthetic human insulin became available, pig insulin was an effective clinical treatment of diabetes in human patients, even though the amino acid sequences of pig and human insulin are not identical. Confidence in such similarities provides a method for predicting the structures of proteins from known structures of close relatives, a procedure known as "homology modelling." However, as evolution proceeds, sequences and structures eventually diverge more radically. Figure 4 shows two distantly related proteins, plastocyanin and azurin, in which the region at the right, containing two sheets packed face-to-face, forms a conserved "core" of the structure, whereas the long loopy region at the left has an entirely different conformation in the two structures.

Protein Structure Prediction

Nature has an algorithm which specifies the three-dimensional structure of a protein from its amino acid sequence alone. We ought to be able to discover it. We should then be able to predict the structures of the proteins inherent in the gene sequences in the human and other genomes, and apply them to practical problems such as drug design. Protein structure prediction has proved to be a difficult problem. Many approaches have been taken, and many claims advanced. However, at present there is no computational method that can consistently produce even a qualitatively correct prediction of protein structure from amino acid sequence, unless a close relative is available.

Suppose you were given the amino acid sequence of a new protein, and asked to predict its structure. What might you try to predict? The most complete information that a prediction might provide is a full set of three-dimensional coor-

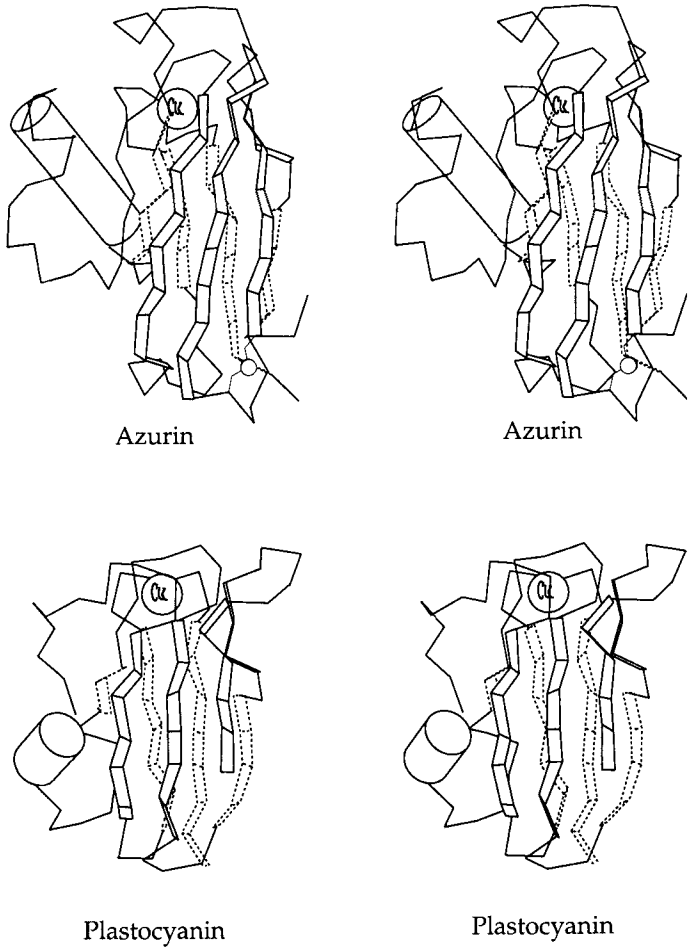

Figure 4. During evolution, gene sequences accumulate mutations, and protein sequences and structures diverge as a result. This figure shows two related electron-transport proteins, poplar leaf plastocyanin and bacterial azurin. The portion of the structures in the right half of the picture, containing the solid and dashed arrays of "ribbon-like" regions—called sheets—and the copper binding site, are well conserved during evolution, whereas the portion of the structure in the left half of the picture has diverged more radically.

dinates for a model of the target protein—a "3-D" prediction. A less ambitious goal would be to predict where in the sequence the regions of helix and sheet appear—a "1-D" prediction. Somewhere in between lie predictions that go beyond "1-D" secondary structure predictions but provide only some qualitative information about the general spatial layout of the folding pattern—let us call these "2-D" predictions.

What kinds of information might you use to predict a protein structure? The ultimate goal is the "pure" *ab initio* approach—use the sequence of the target protein and nothing else. After all, nature does it this way—proteins do not surf the web for databases when they want to fold. But we can, and there has been some success in using information in databanks to identify the fold of a target protein from among the known structures. This is known as the problem of *fold recognition.* Of course it can work only if the structure of one or more proteins with the same fold as the target protein is in your database.

Whom must you satisfy? The following list is sorted, roughly, in decreasing order of difficulty. Most scientists accept that "granting agencies" is the appropriate level to aim at!

Whom must you satisfy?
1. Crystallographers
2. NMR spectroscopists
3. Granting agencies
4. Referees of papers
5. Colleagues
6. Your mother

How in fact could you convince people that you had a successful method for protein structure prediction? Two types of claims are in principle untestable. One is that you can predict the structure of a protein, the structure of which is already known. The other is that you have predicted the structure of a protein, the experimental structure of which is unknown and is likely to remain so for a long time. One must work in the intermediate domain between the known and the not-to-be-known-for-a-long-time, coordinating structure prediction with structure determinations in progress.

To bring order to this activity, to reward those who had made genuine progress, and to refute the claims of those who insisted that they had "solved the protein structure prediction problem," John Moult had the idea of organized blind tests. The idea is that scientists in the process of solving structures will make the amino acid sequence public, but promise to keep the structure secret until an agreed-upon deadline. All who believe that they have a method for protein structure prediction can deposit their predictions before the date of release of the structure. After release, the predictions can be compared with the experiment—to the delight of a few and the chagrin of most. This idea has been developed into the CASP program—Critical Assessment of Structure Prediction—on a biannual basis.

Prediction methods fall into two broad classes, the inductive and the deductive. Inductive methods make direct use of databanks of sequences and structures. Deductive methods are true *ab initio* approaches—the desert island case—attempting to predict protein structure from general principles of physics, chemistry, and biology, without explicit reference to known sequences and structures. Of course the *development* of *ab initio* methods depends on what we have learned from the study of known sequences and structures. The distinction is that the understanding achieved by this study has been distilled into general principles that can be applied without looking up specific information in databases.

Methods for *ab initio* prediction can be divided into two approaches, which I call "Nature" and "Nudger." The Nature approach seeks to understand the natural folding process and then to follow or simulate it. The "Nudger" approach allows any procedure that can force the chain into the proper conformation, even along a path which is not the natural one or is even physically unrealizable.

There is evidence that natural selection has shaped not only the final native states of proteins, but their folding pathways as well. For not only must proteins have evolved to form a stable active conformation, they must achieve that conformation, starting from an unfolded state containing a mixture of random conformations, in a reasonable time. A simple calculation, based on rates of atomic motions in solution, shows that exhaustive exploration of the space of possible conformations would not be fast enough by many orders of magnitude. (This is sometimes called Levinthal's paradox.) There is no evidence that the folding pathway actually affects the final state, although theoretically it could. If alternative folded states were possible, but the pathway evolved to promote one, then we predictors would be *forced* to adopt the Nature rather than the Nudger approach.

Where is the obstacle to structure prediction? We think that we understand the forces that stabilize native protein conformations. It is even possible to write down an explicit conformational energy function of the coordinates. All we need to do is to minimize it. However, it is important to recognize that proteins are, in thermodynamic terms, only marginally stable. In fact the conformational energy of a folded protein is a very small difference between very large opposing terms, a numerical analyst's nightmare.

Is the difficulty that we can't write down the energy function accurately enough, or is the function so complicated that we can't optimize it? One test is to minimize the conformational energies of proteins, *starting from their known native states*. Such calculations do converge to minimum-energy conformations close to the starting point, showing that the energy functions are adequate in the vicinity of the right answer. (Not too surprising, because the functions are defined by fitting parameters to reproduce observed native states.) But this is not enough. A function correct in the vicinity of the minimum will not necessarily provide a complete set of trajectories in conformation space that can enable a program to find the global minimum from an arbitrary starting point.

There are two problems. First, many of the forces that stabilize proteins are short-range. Even if we knew the energy function exactly, if we started a minimization from a random extended, non-compact, conformation, we would find that there are no long-range forces driving the system towards the correct structure. Second, even if we achieved collapse to a compact state, the landscape of the energy as a function of coordinates contains many local minima, separated by high barriers. Many of these local minima will be candidates for the native state. Real proteins overcome these problems by a combination of (1) extensive "parallel processing" in which all the residues *simultaneously* explore their own local dimensions of conformation space, and (2) evolving folding pathways that channel the system towards the right answer. Our computers cannot achieve the parallel processing, our energy functions cannot account for the long-range folding pathways, and our algorithms cannot easily find the global minimum of a complicated multivariate nonlinear function. (Where is the harmonic oscillator when we need it?!)

The difficulty of the *a priori* case has led, as we have noted, to development of empirical methods, based on the known sequences and structures. Prediction methods that use databanks include (1) methods for *homology modelling*—prediction of the target structure from a closely related protein of known structure; and (2) methods for *fold recognition*—assessing the compatibility of the amino acid sequence with the library of known protein folding patterns. These methods are growing more powerful, partly but not entirely because of the growth in the databanks. The more sequences and structures that are known, the more likely that a new protein will be similar to one that is already known. In contrast, *ab initio* methods are improving more sluggishly. A grudging comment about them after a recent CASP competition was that at least ". . . failure can no longer be guaranteed [8]." A pessimist might predict that the growth of databanks will mean that information-based methods will provide pragmatic solutions to

AUTHOR

ARTHUR M. LESK

Department of Haematology
Cambridge Institute for Medical Research
University of Cambridge
Cambridge CB2 2QH
U.K.
e-mail: aml2@mrc-lmb.cam.ac.uk

Arthur Lesk received his Ph.D. in Physics and Physical Chemistry from Princeton University in 1966. He became a Professor of Chemistry at Fairleigh Dickinson University in New Jersey. Since 1977 he has worked at the MRC Laboratory of Molecular Biology in Cambridge, England, where he is now also on the faculty of the University's School of Clinical Medicine. He was a founder member of the Biocomputing Programme at the European Molecular Biology Laboratory in Heidelberg. His education and career choices have been governed by the beliefs that to understand biology you must understand chemistry; to understand chemistry you must understand physics; to understand physics you must understand mathematics; so you might as well start by learning mathematics and then work your way along the list.

such a large majority of questions, that interest in and support for the development of *ab initio* methods will wane. It would be a shame if one of the most interesting of biological computations were thereby lost to computational biology!

I thank the organizers of the Newton Institute Programme, P. Donnelly, W. Fitch, and N. Goldman, for the opportunity to participate in the project; Prof. H. K. Moffatt, N. Goldman, and L. Lo Conte for comments on the manuscript; and the Wellcome Trust for support.

REFERENCES

[1] Wigner, E.P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics* 13, 1–14.

[2] Sankoff, D. and Kruskal, J.B., eds. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, Mass.

[3] Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

[4] Fitch, W.M. and Smith, T.F. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* 80, 1382–1386.

[5] Golub, G. and van Loan, C., *Matrix Computations.* Johns Hopkins Press. Baltimore, 2nd ed. 1989.

[6] Young, G. and Householder, A.S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3, 19–22. (For background and history see [7].)

[7] Blumenthal, L.M. (1938). *Distance Geometries. A study of the development of abstract metrics.* University of Missouri Studies 13, #2.

[8] *New York Times*, March 25, 1997.

Ambigrams

Burkard Polster
Department of Pure Mathematics
University of Adelaide
Australia