$\theta = +12$ K, suggesting dominant ferromagnetic interactions (J. S. Miller, D. M. O'Hare, A. Chackraborty, A. J. Epstein, unpublished results).
39. O. Kahn, *Angew. Chem. Int. Ed.* **24**, 834 (1985).
40. D. Gattesche, *NATO Adv. Study Ser.* in press; A. Caneschi, D. Gattesche, J. Laugier, P. Rey, *J. Am. Chem. Soc.* **109**, 2191 (1987); H. Iwamura, T. Sugawara, K. Itoh, K. Takai, *Mol. Cryst. Liq. Cryst.* **125**, 251 (1985). H. Iwarmura, *Pure Appl. Chem.* **58**, 187 (1986).
41. R. Breslow, *Mol. Cryst. Liq. Cryst.* **125**, 261 (1985); R. Breslow, P. Maslak, J. S. Thomaides, *J. Am. Chem. Soc.* **106**, 6453 (1984); T. J. LePage and R. Breslow *ibid.* **109**, 6412 (1987).
42. This work suggests that a stable triplet may not be necessary for an organic ferromagnet; a stable doublet with a virtually accessible triplet capable of admixing with ground state, as observed for [Fe^{III}(C$_5$Me$_5$)$_2$]$^{\cdot+}$[TCNE]$^{\cdot-}$, should suffice.
43. T. Fukunaga, *J. Am. Chem. Soc.* **98**, 610 (1976); ———, M. D. Gordon, P. J. Krusic, *ibid.*, p. 611.
44. A. J. Hubert, *J. Chem. Soc. C* **1967**, 13 (1967).

45. A.J.E. and J.S.M. gratefully acknowledge partial support by the Department of Energy Division of Materials Science (grant DE-FG02-86ER45271.A000). W.M.R. gratefully acknowledges support by the NSF DMR Solid State Chemistry Program grant 8313710. We thank our co-workers (R. W. Bigelow, J. C. Calabrese, G. A. Candela, S. Chittipeddi, A. Chackraborty, K. R. Cromack, D. A. Dixon, P. J. Krusic, D. M. O'Hare, W. M. Reiff, M. J. Rice, H. Rommelmann, L. Swartzendruber, C. Vazquez, M. D. Ward, D. Wipf, and J. H. Zhang) for the important contributions they have made to this work.

Computers in Molecular Biology: Current Applications and Emerging Trends

CHARLES DELISI

The rate of generation of molecular sequence data is forcing the use of computers as a central tool in molecular biology. Current use of computers is limited largely to data management and sequence comparisons, but rapid growth in the volume of data is generating pressure for the development of high-speed analytical methods for deciphering the codes connecting nucleotide sequence with protein structure and function.

ALMOST 20 MILLION DNA NUCLEOTIDES FROM HUNDREDS of organisms have now been sequenced, and the number continues to rise nearly exponentially, with a doubling time in the range of 2 to 3 years (Fig. 1). For some of the simpler organisms such as *Escherichia coli*, the complete genome will very likely be worked out within the next few years; for humans, the sequence could be available by the turn of the century (*1*). The key to rapid progress would then change from the generation of data to its analysis, from finding the text to reading it.

The development of methods for generating this information has, however, far outstripped the development of methods that would aid in its management and speed its assimilation. As a result, we are witnessing enormous growth in data of the most fundamental and important kind in biology, while at the same time we lack the ability to assimilate these data at a rate commensurate with the potential impact they could have on science and society. Projections based on near-term trends indicate that the problem could rapidly grow worse: during the next 5 to 8 years, projects to map and sequence the human genome (*2, 3*) are expected to increase data flow to about 10^6 bases per day—nearly two orders of magnitude more than the current rate.

The management problem (collecting, organizing, standardizing, disseminating, and so forth) has been widely discussed during the past year, and it will likely be solved during the next several years, largely by scientific cooperation (*4*). The problem of analysis requires linking nucleic acid sequences to the expression, structure, and function of proteins far more rapidly than is currently possible, and will require understanding the series of codes that connect sequence with function far better than we now do. The crucial messages are embedded in the local geometry of the DNA regions that regulate the magnitude and timing of gene expression and in the linear amino acid sequence of the protein itself, which, in a given environment, determines higher order structure and therefore function.

The Computer as Catalyst

As a simple example of converging lines of research in mathematical and molecular biology, consider the implications of data generation for understanding genetic disease. Genes associated with several hundred human diseases have now been assigned to chromosomes (*5*) and often to particular regions, but precise localization has been difficult. Almost 2 years have passed since the cystic fibrosis gene—carried by approximately 1 in 20 Americans of European ancestry—was localized to within a megabase on chromosome 7, but the gene itself still has not been identified. The precise map of such regions is essential for understanding the molecular basis of disease, determining, for example, whether the defect is in the gene itself or in its regulation.

The efficiency of locating genes can be increased by developing a cohesive set of biological, engineering, and mathematical tools. These would consist of an ordered set of DNA clones to provide material for sequencing and mapping, rapid sequencing methods to provide the data for analysis, and new computational methods for the analysis itself (*3*).

The role of computation can be understood by considering the prospect of sequencing a megabase in a day—several hundredfold

This article was written while the author was at the Office of Health and Environmental Research, U.S. Department of Energy, Washington, DC 20545. His present address is Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, NY 10029.

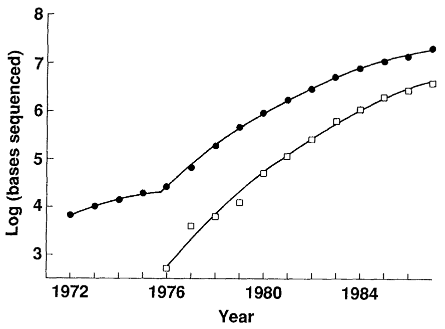


Fig. 1. The number of sequenced nucleotides in GenBank as a function of time (□, human DNA; ●, total DNA). [Courtesy of J. Fickett, Los Alamos National Laboratory]

faster than current technologies permit. A 1-megabase region containing a gene of interest will likely contain 30 to 40 other genes, as well as large regions of noncoding DNA that are present both between and within genes. Neither the boundaries of these regions nor the identity of the gene will be apparent merely by looking at the sequence. Computer algorithms are now being developed, however, that combine artificial intelligence techniques, data on molecular structure, and principles of theoretical chemistry to distinguish genes from intervening sequences and protein coding exons from noncoding introns. With such techniques the sequences of each of the protein products can be deduced more rapidly than by finding and translating the corresponding messenger RNAs. The disease-associated protein can then be sought in the cells most likely to be associated with the disease. The search would be speeded substantially by clues to its most likely cellular location. For example, the gene product for cystic fibrosis is implicated in the regulation of the chloride channel (*6*) and thus could be either a DNA control protein or a regulatory domain of the channel itself. Predictive methods that classify the cellular location and function of protein sequences could narrow the choice of candidate genes. If the structure of the deduced protein sequence were predictable, progress would be even quicker, for then specific tight-binding ligands could be designed and labeled, which would make identification relatively easy.

Computational methods that in effect decipher the messages encoded in DNA and protein sequences differ widely in their state of development. Identification of exon-intron boundaries is about 80 to 90% reliable (*7*). However, genes typically contain several such boundaries, so that the probability of identifying all of them correctly (0.8^n, where $n > 3$) is small. The ability to distinguish integral from peripheral membrane proteins, or to distinguish membrane proteins from all other proteins, is more than 98% reliable (*8*). The ability to predict the three-dimensional structure of a protein is possible under certain circumstances, but not with sufficient accuracy to design a ligand that binds it with high affinity.

This example is highly circumscribed and omits many possibilities. Thus, optimization of mapping strategies (*9*), new approaches to linkage analysis (*10*), molecular dynamics (*11*), and systems dynamics (*12*) are not mentioned, nor will they be discussed here.

Predicting Functional Sites on Nucleic Acids

The regions of interest are the locations of genes and all those sites involved in gene expression: polymerase-binding sites, control protein-binding sites, ribosomal sites, posttranslational modification sites, and so on (*13*). Mathematical identification of such sites is probably the most difficult of the set of problems that is discussed in this article, in part because the higher order structural data that are required for training the learning algorithms are sparse.

Sequence analysis. Several approaches have been taken to the problem of identifying functional regions in sequences. The simplest approach is to look for specific sequences or simple sequence properties that correlate with the function of interest. For example, the dinucleotides GT and AG are well conserved at the 5′ and 3′ termini, respectively, of introns, but, because they occur at many other locations along DNA, their presence alone is a poor predictor of an exon-intron boundary. Even the longer consensus sequences TTGACA and TATAAT, which occur 35 and 20 base pairs, respectively, upstream from messenger RNA start sites in prokaryotes (and thus serve to identify the DNA region recognized by RNA polymerase), are not highly reliable predictors: in the absence of additional information, the probability of correct promoter prediction is about 0.6, since homologous and comparably based sequences will occur at many other locations.

A more subtle property used to distinguish coding from noncoding regions is based on the observation that codons that specify the same amino acid tend to be used with unequal frequencies (*14*). As a result, identical bases in coding regions tend to be in identical codon positions, and this in turn implies the occurrence of sequence periodicities in coding regions that are missing in noncoding regions. The observation can be made quantitative in a number of ways, for example, by calculating the amplitude of the Fourier transform of the appearance of any of the four bases along sequences in coding and noncoding regions.

$$I(\omega) = \left| \sum_{k=1}^{N} (X_i - \overline{X}) \exp(i\omega) \right| \tag{1}$$

where $\omega = 2\pi k/3$. In Eq. 1, X_i is A, G, C, or T and has a value of 1 or 0, according to whether the base at position i is type X. $\overline{X}$ is the number of times base type X appears in the string of N bases, and $I(\omega)$ is a measure of its tendency to repeat with period 3. The amplitude of periodicities in coding and noncoding regions can thus be compared quantitatively and serves as the basis for distinguishing the two regions. Using periodicity (though not Eq. 1), Fickett (*14*) was able to distinguish coding from noncoding regions with 75% reliability, or with 95% reliability if a third "no decision" category was used.

A systematic approach to the problem of recognizing unknown patterns starts with the assemblage of two databases: one with sequences known to have the function of interest, and the other with sequences known not to have the function. The problem of searching a database for common but otherwise unspecified properties has a long history in the field of artificial intelligence (*15*), dating from the development of the so-called perceptron learning algorithm by Rosenblatt in the late 1950s. The method was first adapted to sequence pattern analysis by Stormo *et al.* (*16*) in an attempt to locate ribosomal binding sites in a messenger RNA library containing 124 genes and over 78,000 bases.

The Stormo adaptation, once it has recognized the type of sequence of interest (for example, coding regions), will yield numbers above some threshold when presented with a sequence from the group of interest, and numbers below the threshold otherwise. The method often discriminates perfectly between the two classes of sequences, functional and nonfunctional, on which it is trained, but it is less reliable when applied to sequences outside the training set.

A generalized approach to the problem, which can combine any number of predictive correlates of function, uses the statistical method of discriminant analysis. If attributes are well chosen [$I(\omega)$ and Stormo numbers in the above example], then the range of values they have in the functional class will differ widely from the values they assume in the nonfunctional class. Thus, let $\bar{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a vector of attributes that is distributed according to some probability density function $P(\mathbf{x}/S^+)$ in the functional class, and

$P(\mathbf{x}/S^-)$ in the nonfunctional class, where the parameters in each distribution are fixed by the properties of sequences in the training set. The rule for deciding whether a sequence with attribute vector **x** has the function of interest is as follows: allocate to the functional class if

$$P(S^+/\mathbf{x}) > P(S^-/\mathbf{x}) \qquad (2)$$

and to the nonfunctional class otherwise. In inequality 2, $P(S^+/\mathbf{x})$ and $P(S^-/\mathbf{x})$ are, respectively, the conditional probabilities (obtained by using the Bayes theorem) that a sequence belongs to the functional and nonfunctional classes, given that it has attribute vector **x**. The prior probabilities $P(S^+)$ and $P(S^-)$ are approximated by the fraction of sequences in the database in the functional and nonfunctional classes, respectively.

If discriminant analysis is combined with the perceptron algorithm, exon-intron boundaries can be identified with 80 to 90% reliability, an improvement of 10% over either method alone (*7*). Inclusion of other correlates of function, even if they are not fully orthogonal to the first two, would be expected to increase the accuracy of prediction further.

Both attributes used to predict splice junctions can be quantified with no information other than DNA sequence. The jump from sequence to function, however, explicitly overlooks higher order structure, and the knowledge of such structural details is crucial to understanding function. The regions of interest are three-dimensional structures recognized by other molecules.

Genetic structure. The local geometry of a DNA sequence can be described in a number of ways (*17*). The most complete description is to specify the coordinates of every atom. However, a lower resolution description that models a base as a homogeneous rectangular plank provides a clearer picture of the sources of sequence-dependent variability.

The location of such a plank can be specified by six coordinates: three Cartesian coordinates to locate its center of mass and three Eulerian angles to specify its orientation. Because of chain connectivity, the center of mass of the planks can be located by two coordinates: its radial distance from, for example, the central symmetry axis, and its azimuthal angle about the axis (helical twist) (Fig. 2). If we now consider an axis running lengthwise along the plane of the base pair, two orientational angles can be defined: the inclination of the base axis with respect to a plane perpendicular to the central symmetry axis (base tilt), and the roll of the base pair plane about the axis (that is, the dihedral angle formed by the plane of the planks and a plane perpendicular to the symmetry axis). There remains one additional, internal coordinate arising because the two planks are not rigidly attached to one another. As a result, the planes of a pair need not be parallel but can be twisted relative to one another, like the blades of a propeller.

Average values for base pair coordinates for two different forms of DNA, A and B, are well known, and crystallographic studies have provided new insight into base sequence and solvation-dependent deviations from these averages. Helix twist can be left-handed for certain sequences (*18, 19*). The variations in the values of the coordinates can be substantial—for example, twist ranges from 16° to 44° in A DNA. Such variation and the conformational change accompanying a switch in handedness suggest the geometric basis for variability necessary for selective and specific sequence recognition by DNA binding proteins.

Some progress has been made in relating variations in helical twist, roll, and propeller twist to variations in sequence. Calledine (*20*), in particular, has shown how the ranges that these angles can assume are mutually dependent and are limited by nearest neighbor sequence-dependent steric overlaps. Dickerson (*21*) used this analysis to develop a simple method for predicting the effect of sequence on the variation of roll, propeller twist, and helical twist; a more detailed model has been developed by Tung and Harvey (*22*).

Do these relations between sequence variation and structural variation provide the bridge that is required to more firmly connect sequence to function? Nakata (*23*) indeed found that the Dickerson helical twist angle (but none of the other relations) is an indicator of *E. coli* promoter sequences. However, the relation is weak, indicating that although helical twist does contribute to specificity, it does not dominate recognition. When the twist function is combined with the perceptron method, the classification is correct in ~70% of the cases. This is an improvement over the use of marker sequences alone but is still far from being reliable. However, periodic occurrence of certain dinucleotides produces coupled roll-tilt changes that induce DNA curvature (*24*), which appears to be important in DNA-protein recognition (*25*).

Effective algorithms for identifying functional sites on DNA require a structural database much larger than is currently available, not only to uncover the structural regularities that distinguish various functional sites, but also to develop reliable potential functions that are central to accurate structural prediction algorithms. The importance of developing methods to predict or otherwise determine structure quickly and accurately is apparent when one realizes that even if the code connecting structure to function were known, structural determination would be necessary to use it.

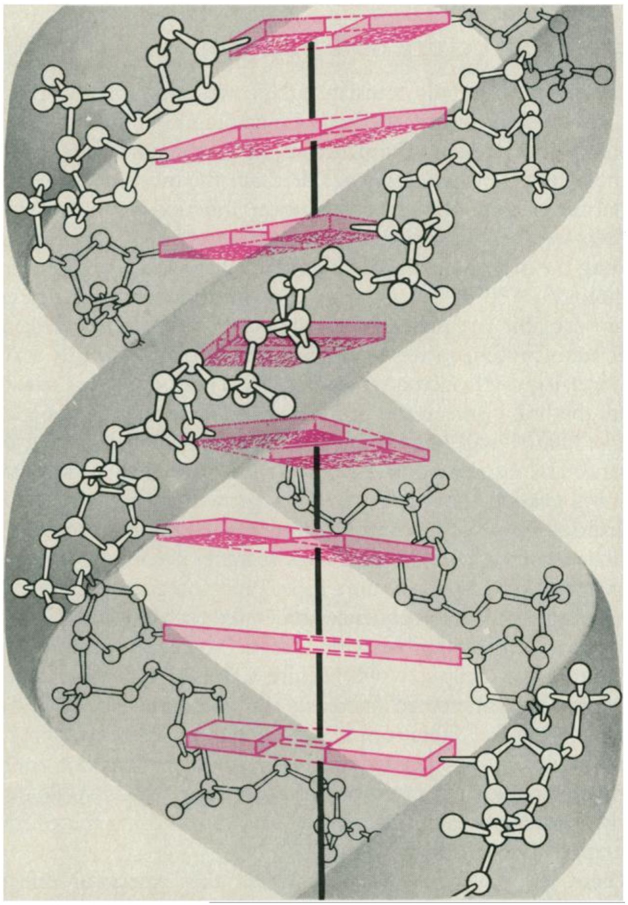


Fig. 2. B form of DNA showing positions and orientations of the base pair planks. The angle formed by the central symmetry (DNA) axis and the long axis of the base pair is called tilt. Roll is rigid rotation of the pair about its long axis, and propeller twist is rotation of one member of the pair relative to the other, also about the pair's long axis. [Illustration copyright by Irving Geis]

The obstacles to accurately calculating nucleic acid structure are somewhat different from those encountered in protein structure prediction (discussed below). Different classes of proteins tend to fold in entirely different ways, and solvent plays a major role in driving folding. In addition, if the initial protein configuration is far from that of the native structure, trapping in local free energy minima is a serious problem. For nucleic acids, the average architecture (that is, average base angles) over regions that are not too large is known. The problem is therefore analogous to predicting the detailed structure of a protein, given the structure of a closely homologous protein. This problem of essentially refining coordinates can be solved with some success for proteins because an empirically established potential energy function is available. The same is not true of nucleic acids, for which the crystallographic data set is relatively sparse, with more than tenfold fewer structures than for proteins.

A second problem is the role of solvent and ionic strength. Solvation will be particularly important for B DNA in which water polymers appear to make an essential contribution to stability. In addition, sequence-dependent salt-induced transitions between different forms of DNA are well documented and must be predictable even by a minimal theory. The role of solvent and salt is too extensive and technical to be adequately discussed in this article except to note that an approximate though successful theory has recently been developed for computing the salt-dependent part of the free energy (*26*).

Protein Structure and Function

An ability to rapidly translate DNA will undoubtedly stimulate the development of new methods for gaining rapid insight—even if it is only generic—into the structure, function, and cellular location of the deduced protein sequence. Perhaps the most direct computational approach to obtaining hints about function is simply to search a database for homologous sequences of known function (*27*). Among the most dramatic discoveries made in this way is the strong homology (>80%) between the protein encoded by the sarcoma virus oncogene and platelet-derived growth factor, and between the *erb*B transforming protein of avian erythroblastosis virus and the epidermal growth factor receptor (*28*). These discoveries demonstrate the link between the products of transforming genes and the intercellular signals known to play a crucial role in cellular growth control, and also demonstrate that at least one step in transformation is a genetic abnormality in a key growth-control signal.

Fruitful homology searches require a large sequence-function database, but, as the database grows, searches become slower. At the current rate of data generation, most supercomputers require a few minutes to compare all sequences generated each day with GenBank. If sequencing rates double every 2 years, however, within a decade the same supercomputer will require a full 24 hours to make the comparison. Anticipated increases in computer speed and the inherently parallel operations in sequence comparisons will probably have little effect on this estimate. The estimate does not apply to comparisons allowing insertions and deletions, which would be one to two orders of magnitude slower, and would bring the computer saturation time closer to 5 years.

There is no obvious way to avoid this emerging difficulty, although some temporary solutions might be possible, for example, limiting searches to a representative sample of the database. Alternatively, the method of characteristic variables could be more fully developed, or a relatively small number of structural motifs might be sufficient to describe most functional domains (see below) (*29*). In either case, rapid determination of higher order protein structure will be a requirement for rapid progress.

One of the simplest principles connecting sequence to structure, and occasionally to activity, is that the solvent-solute system adopts a configuration that minimizes the free energy of the interface. Examples are numerous: large globular proteins in a polar solvent fold so that polar groups generally contact the solvent and apolar groups do not (*30*); membrane-spanning channels form so that polar groups point inward lining the channel and apolar groups are buried in lipids (*31*); membrane-binding peptides adopt a conformation having one face polar and the opposite apolar (*32*), and membrane receptors are anchored by sequences that are predominantly hydrophobic (*33*).

Although the principle is clear, casting it in a form that permits quantitative prediction is difficult. Perhaps the greatest success has been the prediction of membrane-buried sequences (*33*) which can be carried out in a statistically precise manner by discriminant analysis (38). The method assigns protein segments known to be membrane-associated to either a peripheral or an integral category and calculates the odds of correct allocation. The odds function can itself be used in the more general problem of predicting whether a deduced protein sequence, about which nothing else is known, contains an internal membrane segment. The main ambiguity will be in distinguishing segments that are interior to lipids from those that are interior to proteins—both tend to be hidden from polar groups and hence, on energetic grounds, both are expected to be hydrophobic. In fact, if one uses the odds function, the distinction can be made with better than 95% reliability with the longer lengths of the membrane segments being the main parameter distinguishing them from segments buried in globular proteins.

The procedure can be generalized to permit functional classification of a sequence, usually on the basis of just three or four characteristic variables (*34*). Among the more important variables is periodicity in sequence properties. The most general procedure for detecting a dominant periodicity is to fit sinusoids of varying frequencies to the sequence of hydrophobicity values and look for the frequency associated with the best least-squares fit (*35*). A method that gives similar results involves calculating the maximum correlation of the hydrophobicity values with a sinusoid. This is equivalent to obtaining the amplitude of the discrete Fourier transform of the hydrophobicity values, and essentially involves the use of Eq. 1.

Periodicity exerts a major influence on the structure of peptides at interfaces, at the surface of either a protein or a membrane. An example of the former category is the globin family, the general architecture of which consists of eight α-helical segments in a globular arrangement. For globins the dominant characteristic variable is a 3.6-residue repeat in hydrophobicity reflecting a large component of amphipathic α-helical cylinders, that is, α helices with one face predominantly polar (in contact with solvent) and the other predominantly apolar (facing inward) (*36*). On the basis of this and two other variables, globins can be differentiated from every other protein in the Protein Identification Resource (PIR) database with better than 95% accuracy (*34*).

Generic classification is still far from being generally applicable. Between 53 and 64% of the PIR database (depending on whether special proteins are included or omitted) can be allocated to one of 26 functional classes, each of which can be characterized by the joint occurrence of four or fewer characteristic variables. Of the 26 groups, 17 can be filtered from all other proteins in the database with a misclassification error of less than 2%, and the remaining 9 groups can be filtered with errors not exceeding 13% (*34*).

A number of examples exist of simple sequence properties that correlate with activity, such as antigenicity. More specifically, the portions of a protein that stimulate T helper cells—the cells attacked by the AIDS virus, which are central to an effective immune response—can be predicted on the basis of periodic variations in the hydrophobicity of residues along certain portions of the molecule (*37*).

This useful correlate of activity has a simple biophysical explanation. T cells, unlike antibody-producing B cells, are stimulated not by native protein but by portions of a protein after it has been enzymatically digested and presented on the surface of another cell, usually a B cell or a macrophage. The fragment, which in solution is too short to have a well-defined structure, is assumed to be stabilized by interaction with the presenting cell. The simplest model is that the relatively nonspecific reaction with the presenting cell is hydrophobically driven, which suggests a fragment structure with a hydrophobic face in contact with the presenting cell, and opposing polar residues available for the relatively specific reaction with the T cell receptor.

If the hypothesis is correct, periodic variation in the hydrophobicity of residues of antigenic sites would occur with a frequency characteristic of known regular structures: 100° for α structures, 180° for β structures, and 135° for three to ten helices. Of 23 sites known to be antigenic for helper T cells, 18 were found to be amphipathic α helices (*38*), with the probability of chance coincidence being less than 1%. This simple structural correlate of activity is applicable to both the detection and design of T cell antigenic sites, and to a much broader class of peptides including lipoproteins, toxins, and a number of hormones (*32*).

Although the amphipathic principle for membrane-binding peptides appears to be reasonably general, T cell antigenicity may be applicable to between 60 and 90% of immunodominant sequences. The development of a more general procedure will, as with all the other lines of investigation discussed in this article, depend on accurate three-dimensional structure determination—in this case, of the T cell receptor and the major histocompatibility complex molecule on the presenting cell.

Calculating the structure of a large molecule, given only the amino acid sequence, is difficult because of the high probability that the folding pathway generated by free energy minimization will intersect stable but incorrectly folded structures. The likelihood of trapping in a local minimum will be reduced, however, the closer the starting configuration is to the correct structure.

A trial configuration can be obtained in a number of ways; for example, by using high-field nuclear Overhauser effect (NOE) magnetic resonance to place constraints on interatomic distances (*39*) or by using coordinates of homologous regions in proteins whose structures are known (*40*). Analysis of the NOE data can, in principle, provide the complete structure of a peptide backbone at an average resolution of about 3 Å for molecules with masses less than 10^4 daltons. Such measurements can be combined with free energy calculations to determine side chain configurations. Secondary structure can be well characterized by the NOE and can sometimes sufficiently constrain the possible three-dimensional configurations so that a reasonably accurate three-dimensional structure can be deduced (*41*).

The use of coordinates of homologous proteins with known structures has a two-decade history and has been applied to sequences with less than 50% homology. A model of lactalbumin was based on the crystal structure of lysozyme (*42*), which is about 35% homologous and a member of the same superfamily (*43*). Although the procedure is important and widely used, it needs further analysis to determine conditions under which it will provide a structure to within some specified range of accuracy (*44*). Equally important for general applicability is its integration with algorithms characterized by free energy minimization. Beyond that, however,·the extent of its effect on structure determination will be related to the rate at which superfamilies become available, and a representative sample of structures in each superfamily is determined. Although the method will undoubtedly remain important, the prospects for obtaining a sufficiently large number of structures to make homologous extension generally applicable in the near future are not good.

At the present time, the PIR database lists more than 5000 sequences and over 1000 superfamilies. The number of sequences doubles about every 2 years; the superfamily doubling time is closer to 3 years. However, the rate at which crystal structures are entering Brookhaven National Laboratory database is linear at about 40 structures per year. These trends indicate that the rate of structure determination will lag increasingly far behind the rate at which new superfamilies are emerging. Homologous extension, although useful, cannot substitute for the fundamental understanding of protein folding (*45*), which would lead to new and more rapid computational approaches to structure determination.

Finally, the concept of exon shuffling could considerably simplify the structure-function problem in higher vertebrates. The idea was introduced to explain mosaics of commonly occurring sequence patterns, each of which is supposed to be encoded in a single exon and to represent an independently folding functional domain (*46*). Many modern proteins would thus have been formed by duplication and divergence of such exons, which presumably moved about and joined one another in various combinations during evolution. A number of such relatively ubiquitous domains, 40 to 80 residues in length, have been identified. The crystal structures of globins provide direct support for this hypothesis.

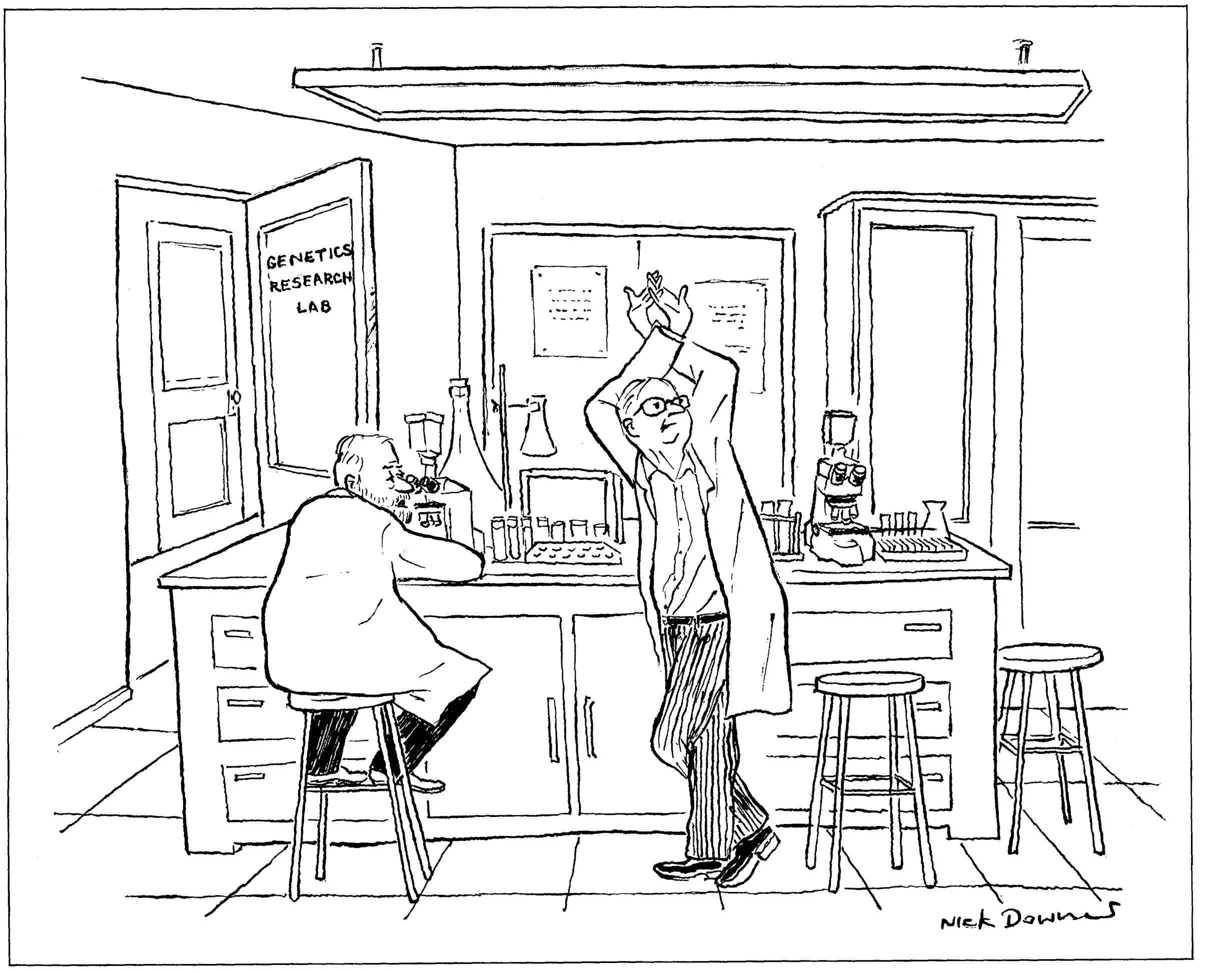
Exon shuffling provides a mechanism for the rapid generation of diversity in terms of relatively few structural motifs. Moreover, to the extent that the hypothesis is valid, the task of determining the structure of a large set of proteins might be reduced to the problem of determining the structure of a smaller number of modular units and understanding the rules by which they pack. Thus, the amount of effort required to determine the structures of a set of proteins consisting on average of *n* modular units would scale approximately as the *n*th root of the number of proteins. The task of understanding function would be similarly reduced.

Where does all this leave us? Data are being generated at an increasingly rapid rate, and the drive to assimilate them will accelerate the development of new approaches for determining molecular structure and function. The computer will also continue to develop at an extraordinary rate, and its potential for an increasingly important role in molecular biology will undoubtedly increase. The full exploitation of computers, however, will only be realized by a large cadre of users with strong backgrounds in mathematics, theoretical chemistry, and molecular biology. Developing the new educational programs that will assure creative use of our cutting edge, high-technology capabilities, is therefore a central challenge as we prepare for the biology of the 21st century.

REFERENCES AND NOTES

1. I assume at least passing familiarity with the elements of molecular biology. Excellent introductions can be found in any number of texts; for example, B. Lewin, *Genes* (Wiley, New York, 1987); J. D. Watson, J. Tooze, D. T. Kurtz, *Recombinant DNA: A Short Course* (Scientific American Books, Freeman, New York, 1983).
2. A. Wada, *Nature (London)* **325**, 771 (1987).
3. *The Human Genome Project* (Health and Environmental Research Advisory Committee Report, Department of Energy, Washington, DC, 1987); *The First Santa Fe Workshop on Sequencing the Human Genome* (Los Alamos National Laboratory Report, Los Alamos, NM, 1986); *Issues Sci. Technol.* **3** (no. 3) (spring 1987).
4. *Future Databases for Molecular Biology* (European Molecular Biology Laboratory–National Institutes of Health Workshop Report, Bethesda, MD, 1987); *Perspectives in Nucleic Acid and Protein Sequencing,* R. Colwell, Ed. (Oxford Univ. Press, Oxford, in press).
5. V. A. McKusick, *Cold Spring Harbor Symp. Quant. Biol.* **51**, 15 (1986).
6. R. A. Frizzell, G. Rechkemmer, R. L. Shoemaker, *Science* **233**, 558 (1986); M. J. Welsh and C. M. Liedtke, *Nature (London)* **322**, 467 (1986).

7. A. Lapedes and R. Farber, *Appl. Neural Net Pattern Recognition Genet. Databases*, in press.
8. P. Klein, M. Kenehisa, C. DeLisi, J. Jacquez, *Biochim. Biophys. Acta* **815**, 468 (1985).
9. L. Goldstein and M. S. Waterman, *Adv. Appl. Math.* **8**, 194 (1987); G. Zehfner and H. Lehrach, *Nucleic Acids Res.* **14**, 335 (1986); J. Cornette and C. DeLisi, *Cell Biophys.*, in press.
10. E. S. Lander and P. Green, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2363 (1987).
11. M. Karplus and J. A. McCammon, *Annu. Rev. Biochem.* **52**, 263 (1983).
12. J. Eisenfeld and C. DeLisi, *Mathematics and Computers in Biomedical Research* (North-Holland, Amsterdam, 1985).
13. C. Burks *et al.*, *Comput. Appl. Biosci.* **1**, 225 (1985).
14. J. W. Fickett, *Nucleic Acids Res.* **10**, 5303 (1982).
15. M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969); J. R. Sampson, *Adaptive Information Processing* (Springer-Verlag, New York, 1976).
16. G. D. Stormo, T. D. Schneider, L. Gold, A. Ehrenfeucht, *Nucleic Acids Res.* **10**, 2997 (1982).
17. Introductions to nucleic acid and protein structure can be found in a number of texts; for example, C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry* (Freeman, San Francisco, 1980), vols. 1 to 3.
18. H. R. Drew and R. E. Dickerson, *J. Mol. Biol.* **151**, 535 (1981); A. Rich, A. Nordhein, A. H. Wang, *Annu. Rev. Biochem.* **53**, 791 (1984).
19. H. Drew *et al.*, *Nature (London)* **286**, 567 (1980); A. H.-J. Wang *et al.*, *ibid.* **282**, 680 (1979).
20. C. R. Calledine, *J. Mol. Biol.* **161**, 343 (1982).
21. R. E. Dickerson, *ibid.* **166**, 419 (1983).
22. C. S. Tung and S. C. Harvey, *Nucleic Acids Res.* **12**, 3343 (1984).
23. K. Nakata, in preparation.
24. L. E. Ulanovsky and E. N. Trifonof, *Nature (London)* **326**, 720 (1987).
25. H. Eisenberg, *Trends Biochem. Sci.* **11**, 350 (1986).
26. D. M. Soumpasis, J. Wiechen, T. M. Jovin, *J. Biomol. Struct. Dyn.* **4**, 535 (1987).
27. W. B. Goad, *Annu. Rev. Biophys. Biophys. Chem.* **15**, 79 (1986); J. B. Kruskal, *Soc. Ind. Appl. Math. Rev.* **25**, 201 (1983); M. S. Waterman, *Nucleic Acids Res.* **14**, 9095 (1986); D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985).
28. R. F. Doolittle *et al.*, *Science* **221**, 275 (1983); M. D. Waterfield *et al.*, *Nature (London)* **304**, 35 (1983); D. Downward *et al.*, *ibid.* **307**, 521 (1984).
29. L. Hood and L. Smith, *Issues Sci. Technol.* **3** (no. 3), 36 (spring 1987).
30. W. Kauzmann, *Adv. Protein Chem.* **16**, 1 (1959).
31. R. M. Stroud and J. Finer-Moore, *Annu. Rev. Cell Biol.* **1**, 317 (1985); J. Finer-Moore and R. M. Stroud, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 155 (1984); H. R. Guy, *Biophys. J.* **45**, 249 (1984).
32. E. T. Kaiser and F. J. Kézdy, *Science* **223**, 249 (1984).
33. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982); D. Eisenberg, *Annu. Rev. Biochem.* **53**, 595 (1984).
34. P. Klein, J. A. Jacquez, C. DeLisi, *Math. Biosci.* **81**, 177 (1986).
35. J. Cornette *et al.*, *J. Mol. Biol.* **195**, 659 (1987). For a review of hydrophobicity, see also G. D. Rose, L. M. Gerasch, T. A. Smith, *Adv. Protein Chem.* **37**, 1 (1985).
36. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
37. C. DeLisi and J. Berzofsky, *ibid.* **82**, 7048 (1985); J. A. Berzofsky, *Science* **229**, 932 (1985).
38. H. Margalit *et al.*, *J. Immunol.* **138**, 2213 (1987).
39. R. Ernst, G. Bodenhausen, A. Wokaun, *Principles of NMR in One and Two Dimensions* (Academic Press, New York, 1987); K. Wuthrich, *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
40. K. A. Palmer, H. A. Scheraga, J. F. Riordan, B. L. Vallee, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1965 (1986); J. Mault and M. N. G. James, *Proteins* **1**, 146 (1986); R. J. Feldman *et al.*, *Ann. N.Y. Acad. Sci.* **439**, 12 (1984).
41. F. E. Cohen, T. J. Richmond, F. M. Richards, *J. Mol. Biol.* **132**, 275 (1979); F. E. Cohen and M. J. Sternberg, *ibid.* **137**, 1379 (1980).
42. D. C. Phillips, *Proc. 7th Int. Congr. Biochem. Tokyo* (1967), p. 63.
43. Proteins are in the same superfamily if the likelihood that the observed homology could have occurred by chance is less than 10^{-3}. For example, immunoglobulin V domains are in a single superfamily and immunoglobulin C region domains are in another. See M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (Georgetown University, Washington, DC, 1976).
44. A start in this direction has been made by A. M. Lesk and C. H. Chothia, *Philos. Trans. R. Soc. London Ser. A* **317**, 345 (1986).
45. A number of reviews of the theory of protein folding have appeared in recent years, including: M. Levitt, *Annu. Rev. Biophys. Bioeng.* **11**, 251 (1982); G. Nemethy and H. A. Scheraga, *Rev. Biophys.* **10**, 239 (1977).
46. C. C. F. Blake, *Nature (London)* **273**, 267 (1978); W. Gilbert, *ibid.* **271**, 501 (1978); R. F. Doolittle, *ibid.* **272**, 581 (1978).
47. I thank G. Bell for helpful suggestions.



"Very good, Michaels—you're a DNA molecule. Now get back to work."