[Go to SCIENCE Online with fewer graphics]

SCIENCE MAGAZINE

HOME | HELP | SEARCH | ARCHIVE | SUBSCRIPTIONS
FEEDBACK | ORDER an article | SIGN IN

MARK GERSTEIN || Change Password || View/Manage User Information || Subscription HELP || Sign Out

Other Articles which have Cited this Article
PubMed Citation || Related Articles in PubMed || Download to Citation Manager

Mapping the Protein Universe

Liisa Holm and Chris Sander

The comparison of the three-dimensional shapes of protein molecules poses a complex algorithmic problem. Its solution provides biologists with computational tools to organize the rapidly growing set of thousands of known protein shapes, to identify new types of protein architecture, and to discover unexpected evolutionary relations, reaching back billions of years, between protein molecules. Protein shape comparison also improves tools for identifying gene functions in genome databases by defining the essential sequence-structure features of a protein family. Finally, an exhaustive all-on-all shape comparison provides a map of physical attractor regions in the abstract shape space of proteins, with implications for the processes of protein folding and evolution.

The authors are in the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton Hall, Cambridge CB10 1SD, UK.

If a living cell is viewed as a biochemical factory, then its main workers are protein molecules, acting as catalysts, transporters, and messengers, among other roles. A human genome encodes about 100,000 of these biological macromolecules. Their functional diversity is made possible by the diversity of three-dimensional (3D) protein shapes, also called "structures," which are capable of highly specific molecular recognition. Understanding or simulating the molecular processes involved in the formation of protein structures and in their biological function is a major challenge of molecular biology. However, in spite of many years of focused research, we still lack a comprehensive and accurate theory of protein structure based on physical and chemical principles. Fortunately, and perhaps unexpectedly, a practical solution to the problem of predicting protein shape and function from amino acid sequence (and thus, ultimately, from nucleotide sequence) is provided by nature itself. Molecular evolution has resulted in a dense network of kinship relations between proteins. By inferring characteristics of one protein based on the function or structure of its relatives, biologists exploit these evolutionary relations to predict protein shape or functional properties.

This exploitation of evolutionary connectivity has become possible because of a wealth of molecular data about proteins from many different species. To date, biologists have read the complete nucleotide (and thus amino acid) sequences of well over 100,000 protein genes (1), and x-ray crystallographers and nuclear magnetic resonance spectroscopists have determined the 3D shapes of several thousand protein molecules (2). The molecular paleontology based on these data reveals a remarkable continuity of molecular evolution. The biochemical function of many proteins has persisted over large evolutionary time scales (even when cellular context has changed, such as in the transition from cells without a nucleus to those with a nucleus). As a protein molecule with an essential functional role evolves in the context of

a living cell, a small number of amino acid residues crucial for its function tend to be strongly conserved (for example, residues that form a catalytically active site), while the rest of the protein sequence eventually undergoes considerable changes. The overall 3D structure also tends to remain essentially unaltered, even when all sequence memory appears to have been lost. This evolutionary resilience of protein 3D structure is the fundamental reason for the importance of protein shape comparison as a computational method in molecular biology.

Recent advances in molecular and structural biology have led to the determination of many 3D protein structures. This article reviews how solving the geometrical shape comparison problem leads to interesting evolutionary observations, to the prediction of function and structure in particular cases, and, on the basis of an all-on-all comparison, to an understanding of the distribution of known structures in shape space.

Comparison by Sequence or by Shape?

Exploiting the observation of evolutionary connections between proteins in order to predict some aspects of structure or function is simple in principle. If a protein is found to be evolutionarily related to another, then information about the function (or shape or enzymatic mechanism, among other attributes) of the one protein can be inferred from that of the other, with varying degrees of accuracy, depending on the evolutionary distance between them. The question then arises as to how evolutionary connections are best detected: by amino acid sequence comparison in 1D or by shape comparison in 3D?

The answer depends on the time interval that has elapsed since the presence of a common ancestor presumed to be similar in function and structure to the two extant descendants. At close evolutionary distances, string comparison between two protein sequences often suffices to establish evolutionary kinship. At larger evolutionary distances, more sophisticated methods must be used to identify subtle similarities in sequence patterns. For example, a method that uses sequence profiles compares probability values for each of 20 amino acids at matching positions in the two proteins under comparison. The most distant relations, however, are no longer detectable by current sequence analysis methods, however sophisticated, and require comparison of the 3D shapes of proteins.

Technically, protein sequence comparison is simpler than shape comparison and is routinely used in studies of protein evolution. Shape comparison can be used only if 3D structures are available (currently in a few percent of all cases), but it is more sophisticated and more powerful than sequence comparison, because similarity of shape remains detectable even though the sequence may have changed beyond recognition in the course of evolution. Comparing protein shapes rather than protein sequences is like using a bigger telescope that looks farther into the universe, and thus farther back in time, opening the door to detecting the most remote and most fascinating evolutionary relations.

An example of what can be done with protein shape comparison is the discovery of a common structural core (a common set of structural elements similarly arranged in space) in two apparently unrelated enzymes from different species with apparently different amino acid sequences. One is mammalian glycogen phosphorylase, a central control point in energy metabolism; the other is a DNA glucosyltransferase that protects the DNA of phage T4 against its own nucleases as it degrades the host's genome (3). Their shape similarity reflects a common chemical mechanism of diphosphate- and sugar-based chemistry, but their substrate specificities and cellular functions, and even their sizes, are very different. Methods for 3D shape comparison were instrumental in this discovery.

Matching 3D Shapes

In geometrical and algorithmic terms, what is involved in shape comparison of two proteins? First of all, a typical 300-residue protein has about 3000 atoms distributed in space according to the convoluted ("folded") trajectory of the polymer chain. Recognizing common substructures between two such structures is in general a very complex combinatorial problem (which points in A are equivalent to which points in B?). The human visual system is very good at recognizing shapes; indeed, classical abstractions of protein architecture (Fig. 1) were established by structural biologists using visual inspection of structures (4, 5). However, as the number of known structures rapidly increased, visual inspection as a general method became inadequate because a human brain cannot easily store the shapes of thousands of complicated macromolecules and cannot easily process the large set of possible substructures. Computers have the advantage of tremendous storage capacity and processing speed but need adequate software. Software development for shape comparison requires (i) a suitable representation of the objects of study, (ii) an objective function to be optimized by (iii) a comparison algorithm, and (iv) appropriate decision rules concerning the significance of the result. Let us look at one way of approaching protein shape comparison.
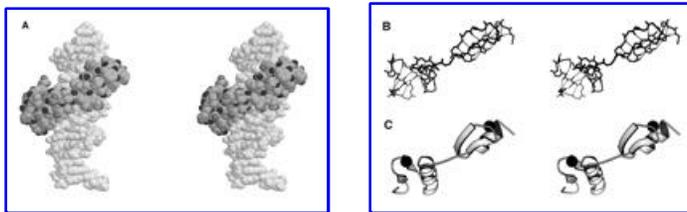

Fig. 1. Protein architecture. The tramtrack protein [Protein Data Bank entry 2drp (30)] is a small protein (525 heavy atoms, 63 residues, and 6 elements of secondary structure), yet it exhibits typical modular protein architecture with two compact structural domains, the so-called zinc fingers. (**A**) The most detailed description of atomic positions is required to understand the function of the tramtrack protein (gray and black, running left to right), which involves binding to a specific base sequence of DNA (white). (**B**) The complicated 3D shape of proteins is encoded in their linear sequence of amino acids. Side chains stripped off, the polypeptide backbone (thick) can be seen meandering from the bottom left to the upper right. Regular patterns of hydrogen bonding (thin lines) between amide and carbonyl groups of the polypeptide backbone give rise to secondary structure (31), shown schematically in (**C**) as arrows for β strands and cylinders for α helices (with zinc atoms as spheres).
[View Larger Versions of these Images (49K GIF FILE/23K GIF file)]

A suitable representation. To make the problem computationally tractable, one first simplifies the representation of protein shapes, keeping essential features. For example, the complicated atomic structure can be represented as a chain trace, that is, the ordered succession of residue centers (C_α atoms) described by their *x, y, z* coordinates (which accounts for about 1 atom out of every 10 in the protein). In these terms, the objective of a comparison of two protein shapes is an assignment of one-to-one equivalence between the C_α atoms, where nonmatching residue centers can be skipped in either chain. In most applications, one also requires that the linear order of equivalent pairs along the sequence is maintained, that is, that the continuity of the polymer chain is considered a key aspect of shape.

An objective function to be optimized. Geometrical objective functions can be formulated in terms of inter- or intramolecular distances yielding, respectively, 3D and 2D comparison problems (Fig. 2). In 3D comparison, one explicitly rotates and translates one molecule relative to the other and measures intermolecular distances between equivalent points in the two chains (Fig. 2A). The objective is to accommodate the largest possible number of equivalent points within small deviations in position,

typically less than 2 to 3 Å. In 2D comparison, 3D shape is described with a matrix of all intramolecular distances between the $C\alpha$ atoms. Such a distance matrix is independent of coordinate frame but contains more than enough information to reconstruct the 3D coordinates, except for overall chirality, by distance geometry methods.
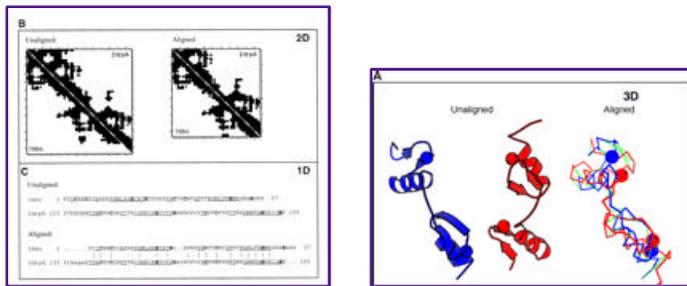
Fig. 2. The meaning of structural equivalence. Shape comparison aims at the one-to-one enumeration of equivalent polymer units in two protein molecules. The problem and solution can be represented in 3D, as a rigid-body superimposition; in 2D, as similar patterns in distance matrices; or in 1D, as an alignment of amino acid sequences. Here, the comparison of the tramtrack protein with another zinc finger protein, the human enhancer-binding protein MBP-1 [Protein Data Base entry 1bbo (32)], is used as an example. (**A**) In the 3D comparison, the problem is to find a translation and rotation of one molecule (red: 1bbo) onto the other (blue: 2drpA). The 3D superimposition (residue centers only, green lines join equivalenced residue centers, zinc atoms as spheres) is not exact because of an internal rotation of the two zinc finger domains relative to one another. (**B**) The 2D distance matrices reveal the conserved structure of the zinc fingers (left: distance matrices of the whole structures; black dots are intramolecular distances less than 12 Å, 1bbo at bottom and 2drpA on top; right: distance matrices brought into register by keeping only rows or columns corresponding to structurally equivalent residues). (**C**) One-dimensional alignment of amino acid strings in the single-letter code; abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. Evolutionary comparison aligns the histidine (H) residues involved in zinc binding (bold; helices and strands of secondary structure are underlined). [View Larger Versions of these Images (34K GIF FILE/23K GIF file)]

A comparison algorithm. How can the 2D matrix comparison be performed? Imagine sliding a (transparent) distance matrix on top of another one. Depending on the register of the two matrices, similar substructures will stand out as submatrices with similar patterns. This view leads to a combinatorial optimization problem of merging matching submatrices to larger consistent blocks of agreement by the removal of intervening rows and columns (Fig. 2B). Algorithmically, this can be treated by a trial-and-error (Monte Carlo) method. In the process of optimization, structurally equivalent regions can be filtered out with a fixed cutoff on acceptable differences of intramolecular distances or, as we prefer, with a continuous function defined in terms of relative distance deviations (6).

Appropriate decision rules. At the end of the optimization process, statistical significance of the comparison score for two proteins can be assessed with empirical criteria (calibrated on a large number of known examples). The results of the shape comparison of two proteins are typically reported in the form of equivalent sets of residues (alignments) (Fig. 2B) or as a 3D view of the matched parts of the two proteins (superimpositions) (Fig. 2A).

Many algorithms have been adapted to the problem of geometrical shape comparison of proteins, including branch-and-bound algorithms, brute force systematic searches, subgraph isomorphism

algorithms, stochastic optimization by Monte Carlo or simulated annealing protocols, genetic algorithms, look-up or hashing methods, dynamic programming, and clustering (7). For most practical purposes, the algorithmic problem of 3D shape comparison of proteins (excepting the problem of comparing protein surface properties independent of the polymer trace) can be considered solved.

Searching 3D Databases

Beyond comparing two proteins, researchers also want to place new protein structures relative to the universe of all protein shapes, or at least relative to all known protein structures. This task is similar to that of finding a match to a fingerprint in a database, but more complicated in that similarities, and not just identities, are of interest. In particular, for a protein structure used as a query, researchers want to see all matches that score above some similarity threshold (for example, such as a threshold defined in terms of statistical significance). Our strategy for efficient searches in the database of 3D structures (2) is to first scan for obvious similarities using fast (but, in general, less accurate) procedures and then to rescan for more subtle similarities using more sophisticated (but slower) algorithms. We turn below to a brief description of these algorithms.

The fast search algorithm achieves simplification and speedup by representing certain repetitive substructures (secondary structure elements, such as α helices and β strands) that consist of perhaps 50 to 150 atoms as 3D vectors anchored at well-defined spatial positions (Fig. 3A). However, this simplification can be misleading when subtle irregularities in the coordinates lead to spurious differences in these vectors for proteins that are actually similar in shape. The algorithm works by storing, in a way convenient for geometrical lookup, a list of spatial relations between such vectors taken from database proteins (8). Here, lookup (or "hashing") is conceptually similar to looking up names in a telephone book. The lookup procedure matches the vector relations taken from the query protein with those in the stored list and proceeds to sample a limited set of spatial superimpositions whenever enough matches are found between the query protein and a database protein. Finally, a dynamic programming step refines these superimpositions and generates detailed residue-level alignments. The search of one structure against the structure database of several thousand structures typically takes only about 5 min on a computer workstation. Other simplified methods achieve similar speed (7). In this way, a large portion (about 90%) of all significant protein-protein shape similarities can be found (Fig. 3A).
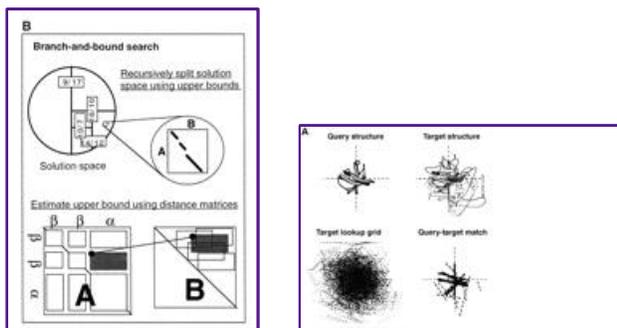
Fig. 3. Two algorithms. (**A**) The 3D lookup is a fast heuristic algorithm that catches easy-to-find structural similarities and is part of the Dali 3D search server (14). The idea is that in favorable cases, 3D superimposition of only a pair of secondary structure elements (SSEs) leads to superimposition of the entire structures. Top: Structure comparison of an SH3 domain of c-Src kinase [1cskA, query structure (33)] with the enzyme papain [1ppn, target structure (34)] reveals similar domain folds, although there is no sequence relation between the proteins and one is much larger. The appropriate orientation of the molecules is found by exhaustive comparison of internal coordinate frames of each protein. An internal coordinate frame is defined by an ordered pair of SSEs (centering one SSE at the origin, aligning it with

the *y* axis, and rotating the molecule around this axis so that the center of a second SSE is in the positive *x-y* plane). Bottom left: Target structure, papain, loaded onto the SSE lookup grid. Each pair of SSEs where the segment midpoints are within 12 Å defines a coordinate frame relative to the grid axes. The figure shows the transformed positions of the 12 SSEs of papain (dotted lines) in each of the ~100 different coordinate frames defined by different pairs of SSEs. Bottom right: The target lookup grid is probed with the SH3 domain, which has four SSEs (thick continuous lines). The coordinate frames shown are the ones yielding the best 3D match of four segments. Iterative extension of a residue-wise alignment starting from the preorientation defined by the SSE match shown here leads to the equivalence of 43 Cα atoms with 1.7 Å root-mean-square positional deviation on an optimal least-squares superimposition. The figure was drawn with MolScript (35). (**B**) A branch-and-bound algorithm (10) is guaranteed to yield the global optimum but may, in the worst case, need an exponential number of steps to do so. An implementation of this algorithm is an essential part of the Dali 3D search server (14). First, protein structures A and B are represented by distance matrices (bottom left and right; each point in a matrix is a residue-residue distance; an internal square is a set of contacts made by two segments; the secondary structure segments are β, β, and α). The problem of shape comparison becomes one of finding a best subset of residues in each matrix (subsets of rows and columns) such that the set of residues in protein A has a similar pattern of intramolecular distances as the set in protein A, as in Fig. 2B. A single solution to the problem is given in terms of the two sets of equivalent residues (an alignment), as shown in Fig. 2C. The solution space consists of all possible placements of residues in protein B relative to the segments of residues of protein A. The key algorithmic idea is to recursively split the solution subspace (schematically shown as a circle at upper left, in which each point is a solution to the problem and the lines divide subsets of solutions) that yields the highest upper bound until there is a single alignment trace left: start with the entire circle; calculate the upper bound for the left (9) and right (17) half; choose the right half and split it into top (upper bound 10) and bottom (upper bound 16) quarters; choose the bottom part and split it (left: 14; right: 12); choose the right part; and so on until the area of solution space has shrunk to a single solution (shown as the residue-residue alignment matrix enlarged at right). The upper bound for each part of the solution space is estimated in terms of a simplified subproblem that asks for the best match of residues in protein B onto a predefined set of residues in protein A (the match is illustrated by the circle-ended line connecting the single square in matrix A with a set of candidate squares in matrix B). The best match is the one with the maximal pair score (sum of similarities of distances between the square in A and the square in B). The predefined set corresponds to residues in secondary structure elements (α, β). The upper bound for each of the segment-segment submatrices of matrix A is found by calculating the similarity scores between the submatrix in A and all accessible submatrices in B. An upper bound of the total similarity score (sum over all segment-segment submatrices in A) for one set of solutions is given by the sum of separately calculated upper bounds for each segment-segment pair of matrix A. The method for choosing constraints that define a set of solutions works in terms of defining allowed residue ranges at each stage of the iteration and is not illustrated.
[View Larger Versions of these Images (38K GIF file/26K GIF file)]

The slower, more sophisticated algorithm is designed to deal with the full combinatorial complexity of comparing two shapes in terms of the spatial trace of the location of residue centers (Cα atoms). As the general problem of finding the global best alignment of two protein traces has the complexity of an NP-hard problem (9), algorithmic solutions must either settle for an approximate solution or risk sifting through an exponentially large search space (approximately N^M possible alignments of a sequence of *N* residues onto a structure consisting of *M* segments of protein trace). To solve this problem to a reasonable approximation, we have adapted the elegant branch-and-bound algorithm by Lathrop and Smith (10) that was originally developed for sequence-structure alignment (to optimally fit the sequence of protein A into the structure of protein B), a problem algorithmically similar to that of distance matrix

comparison. The algorithm iteratively splits the search space of many sequence-segment pairings into subsets, calculates an upper bound of the objective function for each subset, and focuses on further processing (splitting) the subset with the largest upper bound. The chosen series of subsets eventually leads to a subset that contains only a single alignment of protein A with protein B, which corresponds to the exact global optimum of the objective function (Fig. 3B). Continuing the procedure past the global optimum yields suboptimal solutions in monotonically decreasing order. Our adaptation of this branch-and-bound procedure replaces the sequence of protein A by the trace of residue centers of protein A and thus tests all residue-segment pairings─that is, all ways of placing residue centers of protein A at strategically chosen positions in the structure of protein B (atthe beginning of all secondary structure segments, for example).

For reasons of efficiency, we couple this branch-and-bound algorithm to the hierarchical decomposition of a full structure into smaller compact units [similar to "folding unit" decomposition or "domain" decomposition (11)]; that is, we perform the comparison in terms of well-defined substructures. Substructure decomposition is a useful trick (heuristic) because a significant match between two proteins is very likely to contain significant matches between well-chosen substructures. As a result, most placements of residues in protein A onto segments in protein B are pruned before they are examined explicitly. For example, comparing the structures of transducin-α [Protein Data Bank code 1tag, 16 segments (12)] with that of Ras p21 [5p21, 166 residues (13)] leads to a nominal search of 10^{35} spatial arrangements, although the best solution is found after only ~11 s on a fast computer workstation.

The database search methodology containing these two algorithms, plus other tools, is made available over the Internet to users with a coordinate data set describing a 3D protein structure in hand (14). The searches aim to address questions such as which known proteins are related to the query protein in evolution, which parts of a query structure are most conserved, which pairs of proteins have similar internal architecture, and does the query protein represent a new shape (or new fold) not observed to date.

Classifying 3D Shapes

Protein scientists are interested not only in the evolutionary place of particular proteins, but also in a grand view of the architecture of all proteins. Although 10 years ago hand-assembled detailed catalogs of all protein structures would have fit into a review paper or book (5, 15), efficient algorithms of shape comparison and their implementation in computer programs are crucial for coping with the currently more than 4000 structures in the Protein Data Bank (2). Currently, Internet servers rather than printed publications are the preferred medium of dissemination (16). We have recently used shape comparison algorithms to perform an exhaustive all-on-all comparison in order to obtain a quantitative and objective overview of the currently known parts of the protein universe and, if possible, to arrive at a classification of architectural types. In processing the current database, two problems arise, one technical and the other conceptual in nature.

The technical problem is one of redundancy─that is, unequal representation of protein families. For example, there are 230 crystal structures of engineered mutants of phage T4 lysozyme. We can remove the family redundancy by equalizing all proteins with mutual sequence identity greater than 25% (over most of their length, after optimal sequence alignment) because these have essentially complete structural overlap and in most cases similar function (17). Removing such sequence redundancy from the April 1996 release of the Protein Data Bank leaves a set of 740 representative proteins of known structure. Many pairs in this set are still structurally similar to each other, in spite of strong dissimilarity at the sequence level.

Next, in attempting to group structurally similar proteins within the set of 740 representative proteins, there is a conceptual problem, known as the problem of domains. Structural similarities within the set of proteins with unique sequences are typically restricted to only parts of the protein structure. Similar substructures, with relatively sharp boundaries, may recur between several proteins, and conversely, many proteins can be economically described as combinations of recurrent substructures (domains). The notion of such economical description is related to that of minimal encoding in information theory and, in this context, refers to the intuitive goal of defining a small set of large substructures in terms of which most protein structures can be described. In one attempt to achieve this goal, we have combined the notions of compactness and recurrence of domains. A compact domain has minimal surface and maximal interior residue-residue contacts. A recurrent domain is one that appears several times as a recognizably similar substructure in different proteins. This leads to an operational definition of substructures that makes use (i) of the property that normalized distance matrix similarity scores are strongest for complete overlap of large units and (ii) of a physical decomposition of protein structure into a tree of putative folding units at all size levels (18). Given a database of protein shapes, pairwise structural similarities, and alternative decompositions into substructures, the notion of maximal recurrence is implemented by selection of a set of substructures for which the sum of similarities is maximized across the database. As a result, the 740 proteins with unique sequences are split into 1048 domains.

Given this set of domains, one can now group structurally similar domains in a way that was not possible for the set of entire protein structures. There are several options for clustering domains into equivalence groups, none of them, in our opinion, ideal. We chose to group domains similar in shape into "domain fold" classes or simply "fold" classes by a process of average linkage clustering (19). Disregarding small, irregular domains and terminating clustering at an empirically chosen cutoff in similarity, the result is a set of fold classes whose members generally match over all secondary structure elements. This reduces the 1048 domains into 287 structurally unique folds that describe reasonably well the structures of the 740 sequence-unique proteins out of the approximately 4000 known protein structures. The list of currently known fold classes is a good starting point for attempts to better understand the genesis and diversity of protein shapes.

As more and more protein structures are determined experimentally (Fig. 4), automation of the comparison and classification process becomes indispensable. We now continuously monitor the rise in structural knowledge in terms of the appearance of new entries, new protein families, and new fold classes in the Protein Data Bank (2). Simple extrapolation leads us to expect 10,000 database entries, 1600 sequence-unique representative structures (sequence families), and 400 fold classes by the end of 1997. If current trends continue exponentially and without saturation, the 3D coordinates of at least 1 representative of up to 5000 protein sequence families will be known by the year 2000.
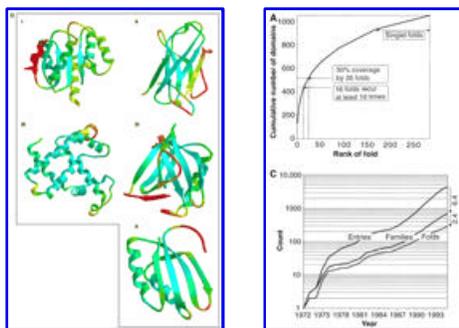

Fig. 4. Recurrent folds. (**A**) A small number of frequently occurring domains (folds) covers a large fraction of all known protein structures. The 287 structurally unique protein domains (folds) are ranked in

descending order of occurrence in the representative set of 740 proteins. Domains ranked 1 through 16 occur 10 or more times each. Domains ranked 1 through 26 cover 50% of all known structures—that is, the essential parts of these structures can be constructed from these domains or described in terms of these domains (within the limits of similarity within a domain class). Domains ranked about 170 or higher occur only once in the current database (singlets). (**B**) Examples of frequently observed fold classes, with one class from each of the attractor regions in Fig. 5 (each attractor region contains several classes, where the term "class" is defined in the text). Color coding indicates which parts of the fold are present in more or fewer members of the class. The color changes from light blue (regions present in 100% of members of the fold class) to red (0% occupancy). The representative classes are defined as follows (attractor, class name, and number of recurrences in sequence-unique set of 740 structures): attractor I: parallel β: COOH-terminal domain of succinyl-CoA synthetase β chain (126); attractor II: β-meander: mouse *opg2* immunoglobulin heavy chain variable domain (52); attractor III: α-helical: myoglobin (15); attractor IV: β-zigzag: COOH-terminal domain of pertussis toxin (14); and attractor V: $\alpha\beta$ meander: COOH-terminal domain of phosphoglycerate dehydrogenase (14). Note that other fold classes in the same attractor region are not shown, but the most frequently occurring are shown in Fig. 5B. (**C**) Growth and redundancy of protein 3D structures in the Protein Data Bank (2). Entry: one of currently more than 4000 sets of protein coordinates in the PDB. Family: collection of proteins set as equivalent if pairwise sequence identity exceeds 25%. Fold: fold class as defined above. The number of new structure entries grows rapidly in time (note logarithmic scale). Redundancy is defined in terms of sequence similarity (sequence families) or structure similarity (fold classes). Currently, there are about 6.4 entries per sequence family and 2.4 families per fold class, for a total of 15 entries per fold. One may expect that in the near future a new fold will appear for about every 15 new entries. The curve of new folds lags behind the curve of sequence-unique families, which indicates the increasing frequency of recurrent folds in newly solved structures (although this may be the result of bias in experimental work). There is no indication that the growth in new fold classes is slowing down at present.

[View Larger Versions of these Images (37K GIF file/31K GIF file)]

Attractors in Shape Space

Conceptually, each protein structure may be imagined as a point in an abstract, high-dimensional fold space. At close range in this fold space, clusters represent protein families related through strong functional constraints (for example, hemoglobin and myoglobin). At intermediate range, clusters are related by shape similarity that does not necessarily reflect similarity of biological function [for example, globins and colicin A (20)]. At long range, the overall distribution of folds is dominated by five densely populated regions, which we call attractors (Fig. 5). Although the current distribution of folds is the result of several effects, including database bias, we put forward the hypothesis that these attractors represent both dominant folding pathways and evolutionary sinks that are the result of physical constraints.
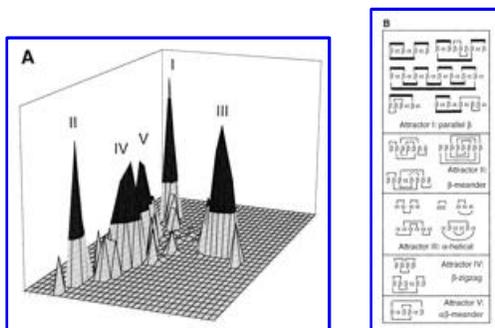
Fig. 5. Fold space attractors. (**A**) Quantification of the pairwise structural similarities in an all-on-all comparison of protein structures allows one to position each structure relative to the others in an abstract, high-dimensional fold space (shape space). The height of the peaks reflects population density (of folds in fold space). The horizontal axes are the two dominant eigenvalues ([21]), and the vertical axis represents the number of protein shapes per unit area (logarithmic scale, arbitrary units). The long-range distribution of different architectures is revealed in a projection down onto the plane based on multivariate scaling, so that proximity in the plot corresponds to correlated structural neighborhoods. (**B**) Forty percent of all known domains (protein substructures) are covered by 16 fold classes (shown as topology diagrams; α, α helix segment; β, β strand segment; thick bar, parallel chain connection between segments; thin bars, antiparallel connection; arc, α helices crossing at roughly right angles). Although each fold class has individual features, most fold classes map to five attractor regions (peaks I through V). All folds with sheets of mainly parallel β strands map to attractor I. The parallel β folds contain a $\beta x \beta$ unit, where the intervening segment (x) is required to reverse chain direction so that the strands are parallel. The $\beta \alpha \beta$ unit has a preferred handedness determined by polymer physics and the natural twist of β strands. Attractor II contains a variety of helical folds. The connectivity of elements in the folds of attractors III and IV contains meander motifs suggestive of the collapse of a long hairpin, either of β strands only or of β strands alternating with a helical pair, $(\beta \alpha \beta)_2$ ([36]). The β zigzag motif of attractor V is simply a series of antiparallel hairpin connections between sequentially adjacent strands. Elementary polymer physics indicates that interactions in space between regions of the chain that are close in sequence are much more probable than those between sequence-distant regions. The β zigzag motif occurs both in flat sheets and barrels, and there is considerable variation in the length of strands (about 4 residues in propeller blades, about 13 in porin barrels). Fold classes other than the most populated 16 are not shown but are accessible from the Dali service over the Internet ([16]).
[View Larger Versions of these Images (68K GIF file/25K GIF file)]

Which basis set represents fold space? We have adopted a multivariate scaling method that discerns the presence or absence of similar features and mathematically amounts to solving an eigenvalue problem ([21]). The method is related to (but different in detail from) principal component analysis and has been used, for example, in archaeology, to arrange sites ("individuals") in a time series based on trends in the composition of excavated artifacts ("attributes") and, in molecular biology, to analyze the correlation between codon usage and level of gene expression ([22]). In our application here, the features are the similarity of an "individual" structure to each other "attribute" structure. We plotted the points in fold space in the 2D plane of the two dominant eigenvectors (Fig. 5A).

What is the evidence for the attractor hypothesis? The five dominant peaks in the distribution of domains in the 2D projection of shape space (Fig. 5A) contain domains with similar secondary structure composition and characteristic topological motifs (secondary structure elements plus loop connections). In the folded structures, the shared motifs are not exposed to solvent, so they are likely to form early on in the folding process and may represent nucleation sites. If this is true, the diversity of different folds that contain these core motifs would represent alternative evolutionary extension paths to add on more elements ([23]). Selective pressure in evolution from random or partially random sequences would be more likely to result in specifically folded stable structures in one of these regions. Economy of description, which underlies our quantitative derivation, may reflect economy of construction.

It would be tempting to speculate that five attractors exhaust the particularly simple pathways of collapsing α helix and β strand elements into globular proteins. However, other solutions to folding up proteins do exist and recur between unrelated families. One such example is the so-called β trefoil fold, which has internal threefold symmetry; it is described as a cone-shaped barrel covered by three β hairpins

(24) and is not in any of the attractor regions. In addition, about 10% of the known fold classes map to small clusters that lack similarity to others. How many more attractor regions are there? Extrapolating from folds that are known to exist, to folds that can exist, is a challenging problem (25). We do anticipate the emergence of some new basic folding patterns from membrane proteins, few of which are known in structural detail. We would be surprised, however, if the number of attractors more than doubled in the next 5 years.

Discovering Evolutionary Links

As more protein structures are determined, the placement of each new protein in shape space makes a contribution to the completion of the map and can, in special cases, lead to a considerable gain in biological knowledge. As an example, let us examine the steps used in unraveling the evolutionary origins of DNA polymerase β, a DNA repair enzyme. When the structure of DNA polymerase β was solved (26), it turned out to be a structural outlier compared to three other DNA and RNA polymerases of known structure. This outlier role seemed to match its peculiarities in function, being smaller and simpler (its polymerase action is stepwise rather than continuous) than the other polymerases. Both features were put in context by the discovery (27) of a close structural resemblance to kanamycin nucleotidyltransferase, an enzyme conferring antibiotic resistance to bacteria. The catalytic domains of DNA polymerase β and kanamycin nucleotidyltransferase share not only a common substructure, but also a sequence signature pattern that maps to the nucleoside triphosphate binding site in the conserved domains (Fig. 5). Pattern searches in sequence databases led to the identification of five additional families of nucleotidyltransferases that are predicted to contain the same substructure responsible for the nucleotide transfer reaction, which in turn led to the definition of an extended enzyme family. In spite of their relation in structure and basic biochemical function, the biological functions of member enzymes are diverse, ranging from nucleic acid synthesis to the regulation of biosynthetic pathways by nucleotidylation (Fig. 6).
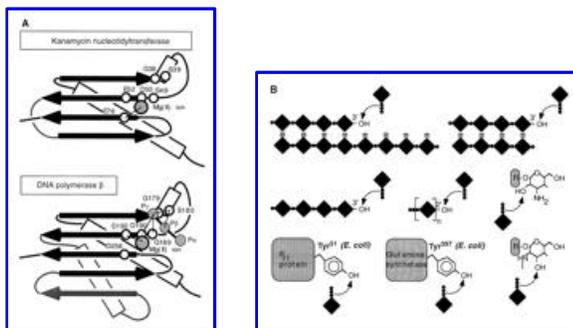
Fig. 6. Evolutionary adaptation of enzyme function. (**A**) Discovery of an essential structure-function feature by shape comparison. A structure database search with DNA polymerase β detects kanamycin nucleotidyltransferase (rather than other known DNA or RNA polymerases) as the nearest neighbor in fold space and reveals conserved residues and structural features supporting the active site. Following up the lead provided by structure database searching with profile searches in sequence databases resulted in the identification of the same characteristics in a large superfamily of nucleotidyltransferases. The biological functions of member families range from DNA repair to regulation of biosynthetic pathways and antibiotic resistance (27). (**B**) Variety of substrate specificity of a common chemical reaction on an essential protein substructure is the remarkable result of biological evolution. All member enzymes of this extended family unified as a result of shape comparison catalyze a common chemical reaction, the coupling of nucleoside triphosphates (black squares and dots) to a free hydroxyl group by means of elimination of pyrophosphate [top row: DNA polymerase β, DNA nucleotidyl exotransferase; middle row: polyadenylate polymerase, $(2'-5')$ oligoadenylate synthetase, kanamycin nucleotidyltransferase; bottom

row: protein P_{II} uridylyltransferase, glutamine synthetase adenylyltransferase, and streptomycin $3'$-adenylyltransferase]. [View Larger Versions of these Images (25K GIF file/27K GIF file)]

Most evolutionary links are identified on the basis of sequence similarity, but the most interesting new discoveries are the result of explorations in the "twilight zone" of sequence similarity. Shape comparison contributes, as it did for DNA polymerase β, by helping to identify subtle but characteristic sequence patterns. The procedure has these steps: structural alignment in 3D of two or more known structures, definition of the pattern of conserved residues in 3D, sequence database searches using that pattern to identify additional candidates, multiple sequence alignment in each candidate family to check consistency of conservation of the search pattern, building explicit 3D models by homology, and verification that the models are physically plausible in terms of sequence-structure fitness (determining how well can the amino acid sequence be accommodated in the 3D structure). This process has already led to the unification of several large sets of functionally related protein families into extended families, with further simplifications expected.

Completing the Protein Map

The growth of sequence and function data from genome projects and 3D structures from experimental structural biology should yield a complete catalog of all proteins soon. Orphan sequences, with no known relatives detectable by sequence alignment, are already diminishing in number (28), and observations of the recurrence of similar substructures in remotely related proteins are more frequent. As more experimentally determined proteins structures become available and computational tools improve, model building by homology will yield a rapidly increasing fraction of all possible 3D models of natural proteins. As a result, the protein folding problem that has traditionally been the focus of computational molecular biology will fade in importance and be replaced by other challenges. In time, computational biologists will move beyond the mere description of evolutionary relations to a quantitative and predictive model of the evolution of proteins. The increasingly complete knowledge of protein structure will be used as a basis for detailed modeling of protein function, protein-protein interactions, and metabolic or signaling pathways. Mapping the protein universe by surveying and classifying protein shapes (29) is a key contribution to these endeavors.

REFERENCES AND NOTES

1. For more information, see the Swiss-Prot database at http://expasy.hcuge.ch/ and the Trembl database at ftp://embl-ebi.ac.uk/pub/databases/treml/ F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977) [Medline]; D. R. Stampf, C. E. Felder, J. L. Sussman, *Nature* **374**, 572 (1995) [Medline]; data sets are accessible at www.pdb.bnl.gov. The Protein Data Bank was founded in 1972 as the global repository for macromolecular structure data.
2. L. Holm and C. **Sander**, *EMBO J.* **14**, 1287 (1995) [Medline]; P. J. Artymiuk, D. W. Rice, A. R. Poirrette, P. Willett, *Nature Struct. Biol.* **2**, 117 (1995) [Medline].
3. L. Pauling and R. B. Corey, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 729 (1951) ; *ibid.*, p. 205; D. B. Wetlaufer, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 697 (1973) [Medline]; C. Chothia, *Annu. Rev. Biochem.* **53**, 537 (1984) [Medline]; ___, A. V. Finkelstein, *ibid.* **59**, 100 (1990) ; B. L. Sibanda and J. M. Thornton, *Methods Enzymol.* **202**, 59 (1991) [Medline].
4. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981) [Medline].
5. Structural similarity S is defined as the sum of similarities of equivalent intramolecular distances such

that

$$S = \sum_{i} \sum_{j} \left(0.2 - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*} \right) e^{-(d_{ij}^*/20\text{Å})^2}$$

where the summation is over all residues *i,j* of the common core, d_{ij}* denotes the arithmetic mean of the Cα-Cα distances d_{ij}A and d_{ij}B in proteins A and B, a relative deviation of 0.2 (20%) is the threshold of similarity, and the exponential factor downweights contributions from pairs at longer distances. The score is elastic, which means that close contacts (for example, adjacent strands in a sheet) may vary 5 ± 1 Å but helix-helix contacts may shift 10 ± 2 Å. The optimal structural alignment is that set of equivalences (*i*A, *i*B) that maximizes *S*.

6. W. R. Taylor and C. A. Orengo, *J. Mol. Biol.* **208**, 1 (1989) [Medline]; A. Sali and T. L. Blundell, *ibid.* **212**, 403 (1990) [Medline]; G. Vriend and C. **Sander**, *Proteins* **11**, 52 (1991) [Medline]; M. T. Barakat and P. M. Dean, *J. Comp. Aided Mol. Design* **5**, 107 (1991); N. N. Alexandrov, K. Takahashi, N. Go, *J. Mol. Biol.* **225**, 5 (1992) [Medline]; D. Fischer, O. Bachar, R. Nussinov, H. Wolfson, *J. Biomol. Struct. Dyn.* **9**, 769 (1992) [Medline]; C. A. Orengo, N. P. Brown, W. T. Taylor, *Proteins* **14**, 139 (1992) [Medline]; R. B. Russell and G. J. Barton, *ibid.*, p. 309; H. M. Grindley, P. J. Artymiuk, D. W. Rice, P. Willett, *J. Mol. Biol.* **229**, 707 (1993) [Medline]; L. Holm and C. **Sander**, *ibid.* **233**, 123 (1993) [Medline]; S. Subbiah, D. V. Laurents, M. Levitt, *Curr. Biol.* **33**, 141 (1993) ; M. S. Johnson, J. P. Overington, Y. Edwards, A. C. W. May, M. A. Rodionov, in *27th Hawaii International Conference on System Sciences:* vol. V*, Biotechnology Computing*, L. Hunter, Ed. (IEEE Computer Society Press, Los Alamitos, CA, 1994), pp. 296--305; K. Diederichs, *Proteins* **23**, 187 (1995) [Medline].

7. L. Holm and C. **Sander**, in *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, C. Rawlings *et al.*, Eds. (AAAI Press, Menlo Park, CA, 1995), pp. 179--187.

8. R. H. Lathrop, *Protein Eng.* **7**, 1059 (1994) [Medline].

9. ___, T. F. Smith, *J. Mol. Biol.* **255**, 641 (1996) [Medline].

10. L. Holm and C. **Sander**, *Proteins* **19**, 256 (1994) [Medline].

11. J. P. Noel, H. E. Hamm, P. B. Sigler, *Nature* **366**, 654 (1993) [Medline].

12. E. F. Pai *et al.*, *ibid.* **341**, 209 (1989) [Medline].

13. The 3D protein coordinates are sent, via e-mail, to dali@embl-heidelberg.de. The Dali server then returns a list of substructures matching all or part of the query structure or asserts that no significant similarity has been found. Precalculated mutual similarities for all known protein structures in the Protein Data Bank (2) are also available from http://www.embl-heidelberg.de/dali and can be viewed as alignments or as 3D views with the use of a Web browser.

14. A. M. Lesk, *Protein Architecture: A Practical Approach* (Oxford Univ. Press, Oxford, 1991).

15. These servers provide Internet access to catalogs of protein 3D structures: Protein Data Bank (http://www.pdb.bnl.gov), Dali (http://www.embl-heidelberg.de/dali), Scop (http://scop.mrc-lmb.cam.ac.uk/scop/), and CATH (http://www.biochem.ucl.ac.uk/bsm/cath/).

16. C. **Sander** and R. Schneider, *Proteins* **9**, 56 (1991) [Medline].

17. The mean and standard deviations of similarity scores were calibrated against pairwise all-on-all comparisons in a database of 220 proteins, as a function of protein size. Shape similarity quantified with the distance matrix comparison scores (6) can then be expressed in terms of normalized *Z* scores—that is, standard deviations above the mean.

18. Average linkage clustering assumes that one knows all pairwise similarity scores and proceeds iteratively by grouping the two most similar domains in the set into a class and representing the similarity relative to this class by an average over the similarities relative to its members.

19. L. Holm and C. **Sander**, *Nature* **361**, 309 (1993) [Medline].

20. M. O. Hill, *Appl. Stat.* **23**, 340 (1974). A multivariate scaling method known as reciprocal averaging or correspondence analysis is a general method for the analysis of contingency tables with m columns and n rows. Here, both rows and columns represent protein structures and the table elements are their pairwise structural similarities (so $m = n$). Let $r_i = \Sigma_j a_{ji}$ be the row totals ($a_{ij} >= 0$). The reciprocal averaging procedure can be represented as the problem of determining a self-consistent set of scores (weights) $\mathbf{x}_i$ from $\mathbf{x}'_i = (\Sigma_j a_{ij} \mathbf{x}_j)/r_i$, where $\mathbf{x}'$ are the new scores and $\mathbf{x}$ are the old scores in an iterative averaging process. A self-consistent set of scores satisfies the eigenvalue problem $\rho \mathbf{x} = (\mathbf{R}^{-1}\mathbf{A})\mathbf{x}$, where $\mathbf{R}$ is a diagonal matrix of the row totals. Following Hill, we iteratively solved for the eigenvectors of the positive semidefinite symmetric matrix $\rho^2(\mathbf{R}^{1/2}\mathbf{x}) = (\mathbf{R}^{-1/2}\mathbf{A}\mathbf{R}^{-1/2})(\mathbf{R}^{-1/2}\mathbf{A}\mathbf{R}^{-1/2})^T(\mathbf{R}^{1/2}\mathbf{x})$, which has a complete set of nonnegative eigenvectors.

21. R. Grantham, C. Gautier, M. Gouy, *Nucleic Acids Res.* **8**, 1893 (1980) [Medline].

22. A. V. Efimov, *J. Mol. Biol.* **245**, 402 (1995) [Medline].

23. S. Onesti, P. Brick, D. M. Blow, *ibid.* **217**, 153 (1991) [Medline].

24. A. V. Finkelstein and B. A. Reva, *Nature* **351**, 497 (1991) [Medline]; A. G. Murzin and A. V. Finkelstein, *J. Mol. Biol.* **204**, 749 (1988) [Medline]; A. M. Lesk, in *Combinatorial Pattern Matching, Proceedings of CPM95*, Z. Galil and E. Ukkonen, Eds. (Springer, Berlin, 1995), pp. 248--260; G. M. Crippen and V. N. Maiorov, *J. Mol. Biol.* **252**, 144 (1995) [Medline].

25. M. Sawaya, H. Pelletier, A. Kumar, S. H. Wilson, J. Kraut, *Science* **264**, 1930 (1994) [Medline]; J. Davies *et al.*, *Cell* **76**, 1123 (1994) [Medline].

26. L. Holm and C. **Sander**, *Trends Biochem. Sci.* **20**, 345 (1995) [Medline].

27. G. Casari, A. de Daruvar, C. **Sander**, R. Schneider, *Trends Genet.* **12**, 244 (1996) .

28. L. Holm and C. **Sander**, *Nucleic Acids Res.* **22**, 3600 (1994) [Medline]; C. Orengo, *Curr. Opin. Struct. Biol.* **4**, 429 (1994) ; A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **247**, 536 (1995) [Medline].

29. L. Fairall, J. W. R. Schwabe, L. Chapman, J. T. Finch, D. Rhodes, *Nature* **366**, 483 (1993) [Medline].

30. W. Kabsch and C. **Sander**, *Biopolymers* **22**, 2577 (1984) .

31. J. G. Omichinski *et al.*, *Biochemistry* **31**, 3907 (1992) [Medline].

32. T. V. Borchert, M. Mathieu, J. Ph. Zeelen, S. A. Courtneidge, *FEBS Lett.* **341**, 79 (1994) [Medline].

33. R. W. Pickersgill, G. W. Harris, E. Garman, *Acta Crystallogr.* **B48**, 59 (1992) .

34. P. Kraulis, *J. Appl. Crystallogr.* **24**, 946 (1991).

35. S. K. Katti, B. A. Katz, H. W. Wyckoff, *J. Mol. Biol.* **205**, 557 (1989) .

PubMed Citation || Related Articles in PubMed || Download to Citation Manager

Other Articles which have Cited this Article

- Gerstein, M., Levitt, M. (1997). A structural census of the current population of protein sequences. Proc. Natl. Acad. Sci. U. S. A. 94: 11911-11916 [Abstract] [Full Text]
- Lander, E. S. (1996). The New Genomics: Global Views of Biology. Science 274: 536-539 [Full Text]
- Xia, D., Yu, C., Kim, H., Xia, J., Kachurin, A. M., Zhang, L., Yu, L., Deisenhofer, J. (1997). Crystal Structure of the Cytochrome bc_1 Complex from Bovine Heart Mitochondria. Science 277: 60-66 [Abstract] [Full Text]

Related Items

FULL TEXT **The Road to Stiff Competition**
Science 2 August 1996; 273 (5275):559 (in Editorial)
F. E. Bloom

SUMMARY **EVOLUTIONARY BIOLOGY: Web-Crawling Up the Tree of Life**
Science 2 August 1996; 273 (5275):568 (in News & Comment)
V. Morell

SUMMARY **TREEBASE: The Roots of Phylogeny**
Science 2 August 1996; 273 (5275):569 (in News & Comment)
V. Morell

FULL TEXT **Computers in Biology**
Science 2 August 1996; 273 (5275):585 (in Computers in Biology; News)
G. J. Chin and T. Appenzeller

SUMMARY FULL TEXT **NETWORKS: Fast Lanes on the Internet**
Science 2 August 1996; 273 (5275):585 (in Computers in Biology; News)
E. Germain

SUMMARY FULL TEXT **NETWORKS: Will Pricing Be the Price of a Faster Internet?**
Science 2 August 1996; 273 (5275):586 (in Computers in Biology; News)
E. Germain

SUMMARY FULL TEXT **COMPUTATIONAL MOLECULAR BIOLOGY: Software Matchmakers Help Make Sense of Sequences**
Science 2 August 1996; 273 (5275):588 (in Computers in Biology; News)
G. Taubes

BIOINFORMATICS: Working the Web With a Virtual Lab and Some Java
Science 2 August 1996; 273 (5275):591 (in Computers in Biology; News)
J. Fischman

BIOINFORMATICS: Do Java Users Live Dangerously?
Science 2 August 1996; 273 (5275):592 (in Computers in Biology; News)
C. Seife

INTERNET: High-Speed Network Will Link Russia's Far-Flung Universities
Science 2 August 1996; 273 (5275):594 (in Computers in Biology; News)
A. Allakhverdov

Four-Dimensional Imaging: Computer Visualization of 3D Movements in Living Specimens
Science 2 August 1996; 273 (5275):603 (in Computers in Biology; Articles)
C. Thomas, P. DeVries, J. Hardin, J. White