Redesigning Life: Organic Chemistry and the Evolving Protein

Steven A. Benner*

Steven A. Benner, born 1954 in New Jersey, USA, received his B.S. and M.S. in Molecular Biophysics and Biochemistry from Yale University in 1976. He finished his PhD. degree in Chemistry at Harvard University under the joint direction of Frank H. Westheimer and Robert B. Woodward. After a sojourn as a Junior Fellow of the Harvard Society of Fellows, he became an Assistant Professor of Chemistry at Harvard University. He joined the Laboratory for Organic Chemistry at the ETH Zürich in October, 1985, as «ausserordentlicher Professor für Bio-organische Chemie».

Techniques from synthetic chemistry, enzymology, and molecular biology are now making it possible to prepare sizeable amounts of proteins whose amino acid sequences have been systematically altered. These techniques permit chemists to apply structural theory to the behavior of biologically important macromolecules using the experimental methods formerly reserved for rather small molecules. This means that there are new experimental tests of hypotheses that connect the structure of proteins to their chemical behavior, their physiological function, and their evolution.

1. Introduction

Structural theory in chemistry, which explains the behavior of matter in terms of its atomic organization, applies to molecules as simple as dihydrogen (formula weight 2) and as complex as a virus (formula weight ca. 10^9). So far, only the first 3 orders of magnitude of its potential scope have been well explored. Nevertheless, research in chemistry over the last 150 years has progressed «upwards». At each step, structural theory has been applied to larger and larger molecules. Three milestones in this progression are shown in Fig. 1.

Macromolecules of biological origin, so-called bio-macromolecules, including proteins (molecular weights typically between 10^4 and 10^6) and nucleic acids (molecular weights typically between 10^5 and 10^7) are the next step in the progression. In contrast to simpler natural products, these bio-macromolecules are linear polymers made from a small number of building blocks. Most proteins are built from 20 amino acids. Most nucleic acids are made from 4 nucleotides. In each case, the backbone of the polymer is repetitive: a repeated amide linkage in proteins, and a repeated phosphodiester linkage in nucleic acids. However, the side chains attached ot the backbone are variable, and the specific sequence of side chains contains information.

The behaviors of bio-macromolecules make them well worth studying. Some are enzymes, potent catalysts that accelerate the rates of chemical reactions by 12 to 14 powers of ten. Others aggregate to form macroscopic structures, like hair, bone, or wood. Some are capable of self-replication, consuming raw material to synthesize copies of themselves. And, aggregates of these molecules, if properly motivated, can move, think, and write review articles.

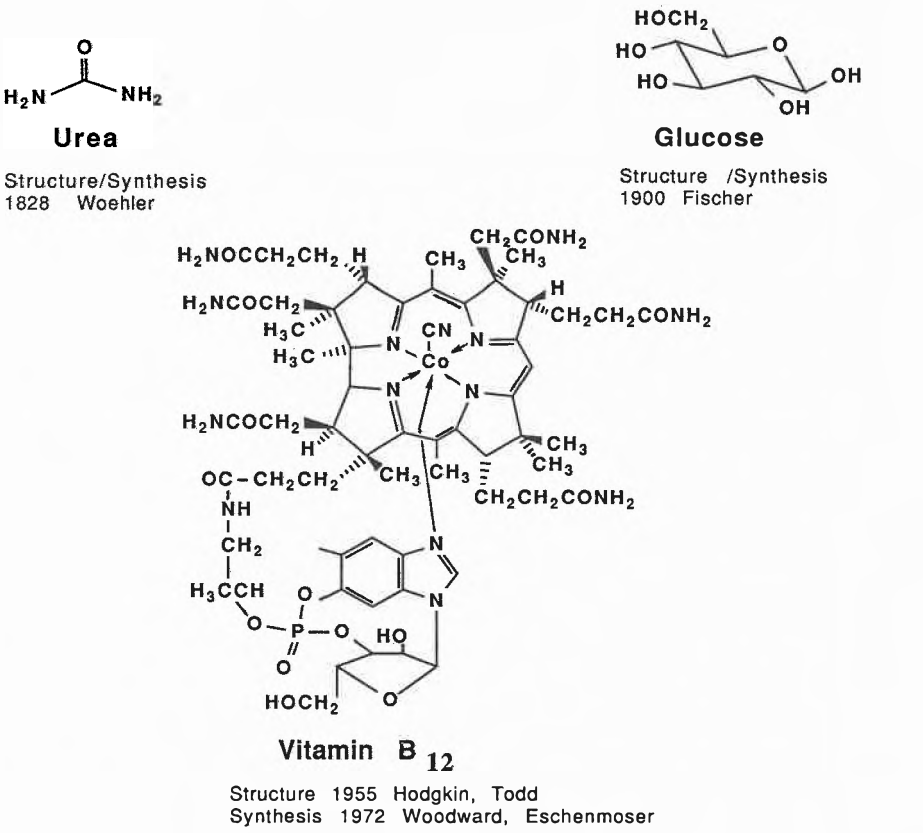

Fig. 1. The structures of three biomolecules whose structural elucidations and syntheses are spaced approximately 70 years apart. Glucose is an order of magnitude more complex than urea, and vitamin B_{12} is an order of magnitude more complex than glucose. However, vitamin B_{12} still has only one tenth the molecular weight of a small protein such as ribonuclease, which itself is two orders of magnitude smaller than a «large» protein. Nevertheless, the progression in organic research, from small molecules to larger ones, continues as the chemist explores the structures of these large macromolecules.

* *Correspondence:* Prof. Dr. S. A. Benner
Laboratorium für Organische Chemie
Eidgenössische Technische Hochschule Zürich
ETH-Zentrum, Universitätstrasse 16
CH-8092 Zürich

2. From Organic to Bio-Organic Chemistry

Organic chemists in general rely on three experimental tools to develop and apply structural theory: (a) purification; (b) structural elucidation at the atomic level of resolution; and (c) synthesis. With these three tools, the chemist applies a simple research method. First, he formulates a hypothesis relating structure and behavior in organic molecules. Then, he synthesizes molecules to test this hypothesis. Finally, he examines the behavior of the synthetic molecules to see if the behaviors match his expectations based on his hypothesis.

Repeated application of this method must expand our understanding of the relationship between structure and behavior in molecules, even for large molecules such as proteins and nucleic acids. However, for the last century and a half, tools for purification, structural analysis, and synthesis have not been sufficiently powerful to meet the chemical challenges presented by biological macromolecules.

This situation has now changed. In the last decade, there has been an explosion in our ability to purify, analyze, and synthesize molecules of this size and complexity. Proteins and nucleic acids can now be purified by high performance liquid chromatography, affinity chromatography, and immunochromatography. Electrophoresis is now routinely applied to seperate nucleic acids that differ in length by a single base. Isoelectric focusing methods allow separation of a protein from contaminating proteins differing in a single amino acid.

Keeping pace with improvements in methods for purifying bio-macromolecules are methods for determining the structure of these macromolecules. The «sequencing» of DNA is now routine. The sequence of amino acids in short polypeptides is now determined automatically by machine. Each decade sees major improvements in the scope of X-ray crystallography as a tool for elucidating the three-dimensional structure of proteins and nucleic acids in the crystal. For structure in solution, spectroscopic methods, most notably two-dimensional NMR, are finally allowing a glimpse at the structure of macromolecules in solution.

Finally, methods for synthesizing biomacromolecules have undergone a revolution. DNA fragments are now synthesized automatically by machine. Synthetic fragments of DNA can be enzymatically joined to synthesize DNA molecules the size of entire genes. Small proteins likewise can be synthesized automatically. RNA can be synthesized enzymatically from DNA.

Molecular biology has provided another set of synthetic tools. Synthetic DNA, cloned in cells, can direct the synthesis of more DNA. Restriction enzymes permit the chemist to cut DNA at specific positions, allowing him to rearrange pieces of DNA. A gene-cloned gene coding for a protein can direct the synthesis, or «expression», of proteins. Further, the structure of the gene can be altered, and the altered gene can be used to make structurally altered proteins.

The tools of chemistry, purification, structural analysis, and synthesis, suitable for handling bio-macromolecules are now in place. In principle, any hypothesis relating structure and behavior in macromolecules of biological origin can be tested by examining synthetic macromolecules that are homogeneous. The potential for advancing our knowledge is great. So are the potential rewards should a detailed understanding of structure and function be developed for bio-macromolecules. To understand these potential rewards, we must understand both the strengths and shortcomings of proteins as potential industrial products.

3. Proteins As a Means to Innovation in Industry?

Proteins, especially those that catalyze reactions, have tremendous potential in the chemical and pharmaceutical industries, as selective catalysts for the synthesis of commercially or biologically important substances, or as commercial products in their own right. Three features of proteins as catalysts are particularly significant:

(1) Enzymes are the products of billions (10^9) of years of biological evolution. Therefore, enzymes they are inherently better than simpler catalysts.

(2) Because, enzymes can effect a specific transformation on a specific compound in a complex mixture, they can participate in designed one-pot multistep pathways that entirely rearrange the carbon skeleton of a precursor, all under equilibrium conditions, just as in a natural biosynthetic pathway. Therefore, enzymes may form the basis of economical, multistep, one-pot synthetic routes to complex chemicals.

(3) Enzymes as macromolecules can have subtly different structures and subtly different physical and catalytic properties as a result. A series of proteins displaying a continuum of properties, ranging from rate and selectivity to stability and solubility, are in principle possible given the appropriate variations in amino acid sequence.

However, the shortcomings of enzymes as catalysts prevents many of these potential advantages from being realized. Central is the problem that only rarely are the enzymes from nature well suited for a specific practical application. Natural enzymes need not have the precise substrate or stereospecificity needed by the biotechnologist. In many cases, natural enzymes are unstable, especially in non-native environments. As a general proposition, we can safely assume that there exists a protein whose structure confers any arbitrary set of properties. But, as present, we do not have a clear idea which structures have which properties. Thus, even with the new tools that permit us to make a protein with any structure, we do not know enough yet to know what structure to make to solve any particular problem.

This contrast, an enormous potential value of applied enzymology obstructed by a lack of specific knowledge about how to manipulate existing enzymes to confer upon them desirable properties, provides an additional motivating force to develop structural theory as applied to bio-macromolecules. In order to take advantage of the special properties of enzymes to make them the catalysts of choice to solve general synthetic problems, we must learn how to perturb the undesirable properties of a protein by deliberately altering its structure.

4. Limitations

In principle, repeated application of an organic chemist's research method that includes hypothesis, synthesis, purification, and characterization will eventually create the knowledge we need to engineer proteins, given sufficient time. However, there are several difficulties that, if not overcome, will make the time required to develop this knowledge very long indeed.

First, the task is immense. The most interesting behaviors in proteins (catalytic power, substrate specificity, stability, biological activity) undoubtedly arise from the interaction of many amino acids in a polypeptide chain. In principle, one can systematically alter many amino acids in a protein to study the dependence of behavior on structure. However, the number of structural variations that must be made in a «brute force» approach is simply astronomical. For a small protein like ribonuclease, there are 2356 single variants that are different from the native structure by one amino acid, 5 505 972 that are different by two, and nearly 13×10^9 that are different by three. One simply cannot make structural variants of a protein randomly in the hope of developing insight into the relation between structure and behavior in this protein.

Many bio-organic chemists have hoped that crystal structures of proteins would contain enough information to allow them to construct a limited number of interesting hypotheses relating structure and behavior that could then be tested using the new tools for handling proteins[1,2]. This hope has so far remained unrealized.

Our understanding of protein chemistry, even given a crystal structure of a protein, is remarkably incomplete. Therefore, hypotheses based on inspection of single crystal structures usually only address very limited issues of protein structure, and generally are applicable only to the specific protein for which the hypothesis is formulated. Examples of commonly asked questions are: Is a specific residue essential for catalysis? Will introduction of a positively

charged amino acid in the active site help bind a negatively charged substrate?[1,2] Often, these questions lead to little more than tests of Coulomb's law or steric hindrance in an active site.

Third, even these limited questions generally cannot be answered by making single mutations, at least not in a way that is interesting to an enzymologist. Mutating an amino acid at the active site may destroy the enzyme's catalytic activity, but the result is uninteresting unless we know whether the residue was directly involved in catalysis, or whether the mutation changed the structure at the active site in a way that influenced the ability of some *other* amino acid to participate in catalysis. It is virtually impossible to know whether the effect of a single structural variation is due to the mechanism envisioned, or by some alternative mechanism.

This is a lesson well appreciated by physical organic chemists familiar with the decades of agony over problems such as the non-classical carbonium ion. Physical organic chemists have long known that «proof» that a particular molecular behavior arises from a specific structural feature can be frustratingly difficult. Bio-organic chemists seem fated to relearn the lesson.

5. More Information

To make the development of structural theory as applied to proteins possible in a reasonable time, more information is necessary to form the basis for structural hypotheses. In our laboratory, we are exploring the possibility that this additional information can come from the recent evolutionary history of proteins. The history can be deduced from the structures and behaviors of many homologous (evolutionarily related) proteins.

The approach is based on the notion that evolutionary processes have been engineering proteins for billions of years, introducing single amino acid changes into existing proteins to obtain new proteins with desired physical and catalytic properties. If this hypothesis is correct, studying how evolution engineers proteins is a strategy for developing insight that will allow the chemist to do the same. We shall first present evidence supporting this notion, and shall then report recent progress in synthesizing, cloning, and expressing genes for ribonuclease (RNase), a protein especially well suited for exploiting «evolutionary guidance» in understanding structure-function relationships in proteins.

6. Does Evolution Engineer the Behavior of Proteins?

Traits of enzymes are either selected or they are not, depending on whether they influence the survival of the host organism[3]. As natural selction is the only mechanism for obtaining functional behavior in living systems, only selected traits reflect function. Non-selected traits reflect history, including both the traits of ancestral enzymes and any constraints that may prevent those ancestral traits from «drifting»[3]. They need not reveal anything fundamental about the workings of biological catalysts.

To use evolutionary history as a guide, we must know which details in the behavior of proteins reflect function achieved by natural selection, and which reflect the random accumulation of «neutral» mutations by a process of «drift»[3].

There are many subtle traits of enzymes that bio-organic chemists study where this question can be asked. For example, dehydrogenases dependent on nicotinamide cofactors transfer only one of the two hydrogen atoms at the 4-position of NADH. About half of the enzymes so far transfer the pro-*R* (or "A") hydrogen, about half transfer the pro-*S* (or "B") hydrogen[4]. Table 1 shows a list of dehydrogenases listed in the order of the catalog number used by the Enzyme Commission to identify the enzyme.

The impression gained from an inspection of Table 1 is that the hydrogen transfered by any particular dehydrogenase is «random». Therefore, this may not be the product of natural selection, and may not be able to provide fundamental information about protein catalysis. However, in the absence of an alternative testable hypothesis that interprets the same data in terms of function, we cannot be sure whether the appearance of «randomness» reflects actual randomness, or whether it simply reflects our ignorance of the underlying functional significance of the trait.

Recently, we proposed such a functional theory that suggests that dehydrogenase stereospecificity can be explained as a trait engineered by selective forces to obtain an enzyme with optimal catalytic efficiency. Our model is based on four hypotheses[5]:

(a) Dehydrogenases have evolved to reflect stereoelectronic principles. The "A" hydrogen is transfered if the enzyme has evolved to bind the cofactor in the *"anti"*-conformation; the "B" hydrogen is transfered if the enzyme binds the cofactor in the *"syn"*-conformation (Fig. 2)[6].

(b) The *"anti"*-conformation of the cofactor is a weaker reducing agent than the *"syn"*-conformation.

(c) Enzymes are catalytically optimal when they bind cofactor and substrates so that the free energy drop between enzyme-substrate and enzyme-product complexes reflects some of the chemical potential drop of the reaction under physiological conditions[7].

(d) Dehydrogenases have evolved to be catalytically optimal.

These hypotheses predict that the stereospecificity of dehydrogenases should correlate with the redox potential of the substrate that the enzyme had evolved to handle. This was in fact found to be the case. Rearranging the enzymes listed in Table 1 so as to order them not by catalog number, but rather by the stability of their presumed natural substrate (Table 2), a correlation between redox potential and stereospecificity is apparent. Reactive carbonyl compounds are reduced by the pro-*R* hydrogen, unreactive carbonyl compounds are reduced by the pro-*S* hydrogen. This correlation is fitted by 125 of the 130 enzymes studied so far[8].

The fact that the apparently random data can be organized by a functional model suggests, but does not prove, that evolutionary selection pressures have se-

Table 1. Stereoselectivity of some dehydrogenases using NADH, listed by Enzyme Commission catalog number. By convention, the "A" and "B" hydrogens are the pro-*R* and pro-*S* hydrogens.

E.C.	Enzyme	Stereochemistry
1.1.1.1	alcohol dehydrogenase (yeast)	A
1.1.1.3	homoserine dehydrogenase	B
1.1.1.6	glycerol dehydrogenase	A
1.1.1.8	glycerol-3-phosphate dehydrogenase	B
1.1.1.26	glyoxylate reductase	A
1.1.1.27	L-lactate dehydrogenase	A
1.1.1.28	D-lactate dehydrogenase	A
1.1.1.29	glycerate dehydrogenase	A
1.1.1.30	3-hydroxybutyrate dehydrogenase	B
1.1.1.35	3-hydroxyacyl-CoA dehydrogenase	B
1.1.1.37	malate dehydrogenase	A
1.1.1.38	malic enzyme	A
1.1.1.40	malic enzyme (NADP)	A
1.1.1.51	β-hydroxysteroid dehydrogenase	B
1.1.1.60	tartronate-semialdehyde reductase	A
1.1.1.62	estradiol 17 β-dehydrogenase	B
1.1.1.64	testosterone β-dehydrogenase	B
1.1.1.72	glycerol dehydrogenase (NADP)	A
1.1.1.79	glyoxylate reductase (NADP)	A
1.1.1.81	hydroxypyruvate reductase	A
1.1.1.82	malate dehydrogenase (NADP)	A
1.1.1.91	aryl alcohol dehydrogenase	B
1.1.1.100	3-oxoacyl acyl carrier protein reductase	B
1.1.1.108	carnitine dehydrogenase	B
1.1.1.50	3α-hydroxysteroid dehydrogenase (P. test.)	B

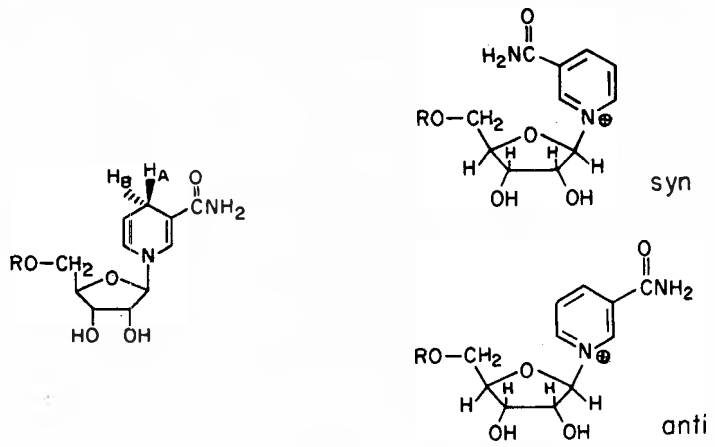

Fig. 2. syn- and anti-conformations of nicotinamide cofactors, and the two distinct hydrogens (A and B, or pro-R and pro-S) at the 4'-position of NADH. Enzymes distinguish between them. Although the chemical distinction is quite subtle, it seems to be large enough for enzymes to have evolved to select one hydrogen atom over the other for mechanistic reasons. This behavior is one of the most subtle to be explained in terms of selected function, where predictions made by the functional model are experimentally verified.

Table 2. Stereoselectivities of dehydrogenases arranged by the equilibrium constant [NADH] [carbonyl] [$H^{\oplus}$]/[$NAD^{\oplus}$] [alcohol] for the redox reaction that the enzyme has evolved to catalyze.

Enzyme	E.C.	lg E_{eq}	Stereochemistry
glyoxylate reductase	1.1.1.26	17.5	pro-*R*
glyoxylate reductase ($NADP^+$)	1.1.1.79	17.5	pro-*R*
tartronate-semialdehyde reductase	1.1.1.60	13.3	pro-*R*
glycerate dehydrogenase	1.1.1.29	13.3	pro-*R*
glycerol-1 dehydrogenase	1.1.1.72	12.8	pro-*R*
hydroxypyruvate reductase ($NADP^+$)	1.1.1.81	12.4	pro-*R*
malate dehydrogenase	1.1.1.37	12.1	pro-*R*
malate dehydrogenase ($NADP^+$)	1.1.1.82	12.1	pro-*R*
malic enzyme	1.1.1.38	12.1	pro-*R*
malic enzyme ($NADP^+$)	1.1.1.40	12.1	pro-*R*
L-lactate dehydrogenase	1.1.1.27	11.6	pro-*R*
D-lactate dehydrogenase	1.1.1.28	11.6	pro-*R*
ethanol dehydrogenase (yeast)	1.1.1.1	11.4	pro-*R*
glycerol-2 dehydrogenase	1.1.1.6	11.3	pro-*R*
glycerol-3-phosphate dehydrogenase	1.1.1.8	11.1	pro-*S*
homoserine dehydrogenase	1.1.1.3	10.9	pro-*S*
carnitine dehydrogenase	1.1.1.108	10.9	pro-*S*
3-hydroxyacyl-CoA dehydrogenase	1.1.1.35	10.5	pro-*S*
3-hydroxybutyrate dehydrogenase	1.1.1.30	8.9	pro-*S*
3β-hydroxysteroid dehydrogenase	1.1.1.50	8.0	pro-*S*
testosterone dehydrogenase	1.1.1.64	7.6	pro-*S*
3-oxoacyl-ACP dehydrogenase	1.1.1.100	7.6	pro-*S*
β-hydroxysteroid dehydrogenase	1.1.1.51	7.6	pro-*S*
estradiol 17β-dehydrogenase	1.1.1.62	7.7	pro-*S*

lected for one stereochemical outcome over the other in different enzymes.

Further evidence that this stereoselectivity reflects selected function rather than random drift can be found by testing predictions that the functional theory makes, especially those that differ from those made by historical models. For example, one generalization regarding dehydrogenase stereoselectivity, known as «Bentley's first rule», is that enzymes from different organisms but acting on the same substrate all have the same stereochemistry[9].

Historical models treat this as evidence that all enzymes acting on the same substrate are related, with the stereospecificity of the ancestral enzyme being conserved. Functional models such as ours argue that this reflects a selection for a particular stereochemistry for functional reasons related to the redox potential of the natural substrate, the same for all of the enzymes in question. If the functional theory is correct, we would expect that enzymes acting on substrates having redox potential at the break in the correlation would not be under strong selective pressure to favor one stereospecificity over the other. Therefore, we might expect that ethanol dehydrogenases, for example, to come in two stereochemical varieties if the functional model is correct, but not if a historical model is correct.

Therefore, we have examined a number of ethanol dehydrogenases, and find that indeed, some transfer the pro-*S* hydrogen (the enzyme from *Drosophila melanogaster*) while others transfer the pro-*R* hydrogen (the enzyme from yeast). In contrast, all malate dehydrogenases transfer the pro-*R* hydrogen, including those from mammals, insects, bacteria, archaebacteria, and plants. Likewise, all 3-hydroxyacylcoenzyme A and hydroxybutyrate dehydrogenases examined so far transfer the pro-*S* hydrogen. These data are consistent with our functional model, but not with simple historical models.

Analogous arguments have been constructed and tested for a range of stereochemical behaviors of enzymes, a range of kinetic and thermodynamic details of enzymatic reactions, dynamic behaviors of proteins, and physical properties of proteins. We are now able to make reasonably general statements as to what traits in proteins are selected, and therefore functional, and which are not.

The conclusions of these studies can be simply stated. In general, enzymes have evolved to be highly refined catalysts optimally suited for effecting specific transformations on specific natural substrates under defined conditions. Indeed, if we understand the physiological role of an individual enzyme, we can in many cases predict its properties based on the adaptation of the protein to fill this role.

Thus, it appears as if natural selection is capable of engineering even subtle details of enzymic behavior without major constraints. This in turn suggests that the process by which evolution engineers proteins is likely to be interesting to those seeking insight into how to engineer proteins in the laboratory. If we can learn how evolution engineers proteins by changing one amino acid at a time, we may be able to gain insight that will permit the chemist to do the same.

7. Synthesis, Cloning, and Expression of a Gene for Ribonuclease

To illustrate how «evolutionary guidance» can form the basis of a research program to apply structural theory to macromolecules, let us consider briefly ribonuclease (RNase; see Fig. 3). RNase is a small protein that catalyzes the hydrolysis of ribonucleic acid (RNA). The RNase from the digestive tract of bovoids is composed of 124 amino acids, and has a molecular weight of approximately 14000. A crystal structure at atomic resolution was obtained for a variant of this protein 20 years ago by *Fred Richards*, and seven other crystal structures of variants of the protein have been obtained since.

RNases can be found in the digestive tract of many organisms. Particularly important is digestive RNase in ruminants (artiodactyls), where the protein has undergone a rapid recent evolutionary divergence as ruminants themselves diverged to occupy a wide range of environmental niches. Thanks in large part to the extensive work of *Beintema* et al.[10], we now know the amino acid sequences of over 50 of these divergent RNases. From these

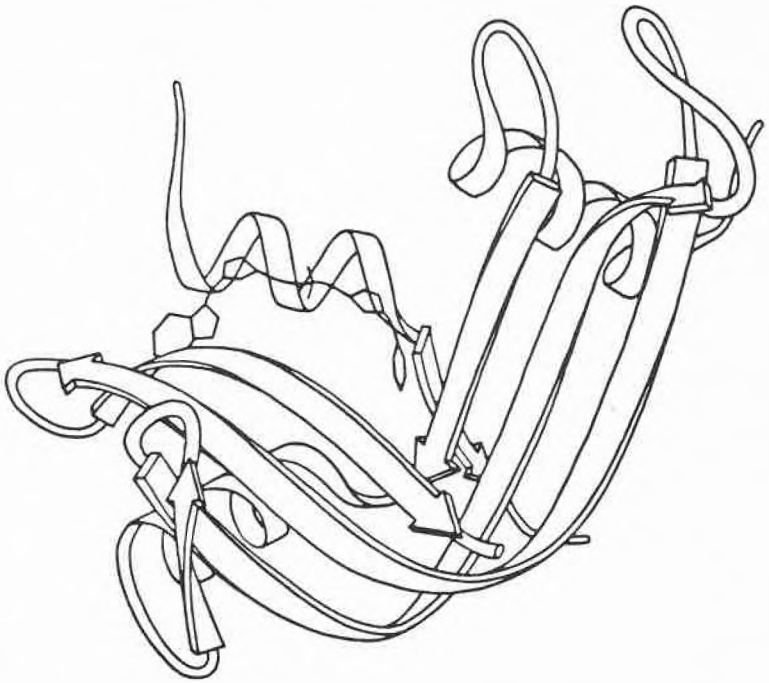

Fig. 3. The folded form of ribonuclease (RNase), showing several alpha-helices and beta-strands. One of the smallest enzymes known, RNase is an ideal target for studying the relationship between structure and behavior in proteins.

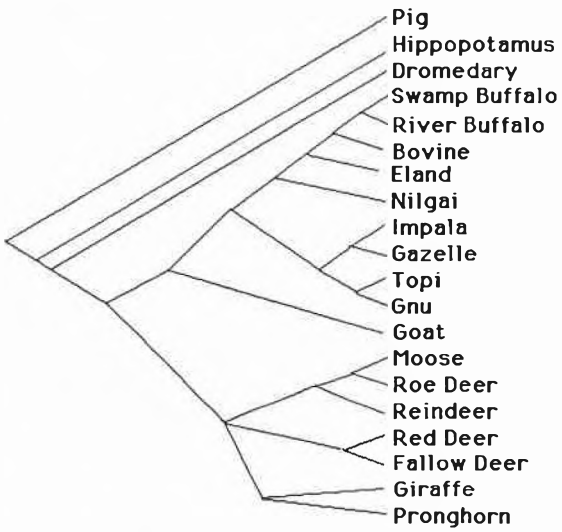

Fig. 4. A hypothetical evolutionary tree connecting organisms whose digestive ribonucleases (RNases) have been sequenced. Branch points in the tree represent ancestral RNases whose sequences can be deduced by extrapolation back from the sequences of modern enzymes. As the behaviors of the modern proteins have diverged considerably, and the sequences have diverged as well, the tree offers a step-by-step path where proteins with substantially different structures and behaviors can be interconverted, one amino acid residue at a time. Thus, evolution, which has engineered the properties of RNases by systematically altering the structure of the protein, can guide the chemist as he attempts to do the same.

modern enzymes, a tree can be constructed (Fig. 4) that represents the divergent evolution of this protein.

RNase has diverged to perform roles in places besides the digestive tract. RNases are found in blood and seminal fluid. In several cases, the sequences of these proteins are known; the proteins are clearly homologous to the RNases from the digestive tract. Further, the proteins have some interesting physical, catalytic, and biological properties. For example, the seminal enzyme is a dimer[11], it acts on double-stranded nucleic acid, and it is a potent inhibitor of tumor growth[12].

Further, *Vallee* and his collaborators recently isolated a protein secreted by tumor cells. The protein's function apparently is to cause the growth of blood vessels to feed the tumor, and hence is called «angiogenin». Angiogenin is also homologous to RNase[13].

Finally, the structures of eosinophil-derived neurotoxin and eosinophil cationic protein have both recently been shown to be homologous to RNase. The proteins, from the human eosinophil granule, are neurotoxic, and cause in rabbits a neurological syndrome characterized by ataxia, muscular weakness, and muscle wasting[14].

The participation of proteins related to RNase in so many biological functions was quite unexpected. In itself, it is noteworthy, as it suggests that RNA molecules are important as short-range messengers mediating complex biological processes important in development and disease. However, the existence of so many different proteins, all related in structure, but different in their physical, catalytic, and biological properties, also offers a tremendous opportunity to chemists seeking to understand how structure and behavior are related in proteins.

The branch points in the tree in Fig. 4 represent RNases from ancestral (now extinct) organisms. Just as the vocabulary of the Indo-European language can be constructed by extrapolation of the vocabularies of the modern languages descendent from it, sequences of these ancestral RNases can be reconstructed by extrapolation from the sequences of modern proteins. Thus, the evolutionary tree provides «paths» connecting structurally different (but related) enzymes with different behaviors. Each path, consisting of a series of proteins differing by a small number of amino acids from adjacent intermediates, is therefore a recipe, or guide, for how to change the behavior of a protein by changing its structure.

There are many different behaviors about which the recent evolutionary history of RNase is informative. In the divergent evolution of the structure of RNase, amino acids in 70% of the positions have been altered. The behavioral divergence is also rather significant. The homologous set of ribonucleases includes some proteins that are monomers and some that are dimers. Some act on single-stranded RNA; others act on double-stranded RNA. Some are quite thermally stable; others are thermally unstable. Some, the angiogenins, apparently promote the growth of tumors; others inhibit the growth of tumors. Some are neurotoxic; others have no neurotoxic activity.

Of course, it is likely that the behavioral divergence between any pair of homologous proteins reflects only a subset of the observed structural divergence. Nevertheless, the paths allow us to form a manageable number of hypotheses about the relation between complex behaviors in proteins and multiple structural changes. These hypotheses can then be tested by synthesizing proteins with structures that are intermediate between the two modern proteins with different behaviors.

For example, variations at positions 25, 31, and 32 in RNase appear to control the quaternary structure of RNases. Variations at positions 3, 16, 17, 19, and 20 appear to control the three-dimensional packing of different domains of the protein. Variations at positions 55, 62, 64, 76, 111, 113, and 115 appear to control parts of the substrate specificity of the protein.

Thus, the evolutionary history of the protein has allowed us to focus on a relatively small number of amino acids as responsible for controlling complicated behaviors in the protein. Proteins with structures altered in these positions then can be synthesized to test hypotheses regarding individual residues or groups of residues.

To implement this approach, several steps were taken. First, we synthesized a gene coding for RNase using a combination of chemical and enzymatic methods[15] (Fig. 5). A synthetic gene has several advantages over a gene cloned from a natural source. Because of the redundancy of the genetic code, a synthetic gene can be designed to contain unique restriction sites at regular intervals throughout the gene. These restriction sites facilitate modification of the gene. To introduce a mutation in a designed gene, a module is cut from the gene by digestion with the restriction endonucleases whose sites flank the region to be replaced. The gene is then religated with a module of synthetic duplex DNA that codes for the desired mutation. This strategy for rapidly creating altered genes is called «modular mutagenesis».

The synthetic gene, consisting of over 700 bases joined in two strands in a precise order, was cloned in *E. coli*. The next challenge was to persuade the bacterium to manufacture RNase from the gene, a process called «expression». This was done by fusing the gene coding for RNase with part of a gene coding for another protein, beta-galactosidase. The hybrid protein is more stable in *E. coli*, and is easily purified by chromatographic methods.

The last step is the release of expressed RNase[16], and refolding the protein to its native structure. RNase so prepared has properties corresponding to those of authentic material. Similar methods are now routinely applied to make RNase proteins with altered structure, via the expression of mutated genes.

```
5'    Ile Glu Gly Arg  Lys Glu Thr Ala Ala Ala Lys Phe Glu Arg
AATTC-ATT-GAG-GGT-CGT==AAA-GAA-ACT-GCG-GCC-GCA-AAA-TTT-GAA-CGT-
3'  G-TAA-CTC-CCA-GCA==TTT-CTT-TGA-CGC-CGG-CGT-TTT-AAA-CTT-GCA-
EcoRI       MnlI                        NotI      DraI
                                        XmaIII

Gln His Met Asp Ser Ser Thr Ser Ala Ala Ser Ser Ser Asn Tyr Cys
CAG-CAT-ATG-GAC-AGT-TCC-ACG-TCC-GCC-GCT-TCT-TCT-TCG-AAT-TAT-TGT-
GTC-GTA-TAC-CTG-TCA-AGG-TGC-AGG-CGG-CGA-AGA-AGA-AGC-TTA-ATA-ACA-
     NdeI                                  MboII   TaqI

Asn Gln Met Met Lys Ser Arg Asn Leu Thr Lys Asp Arg Cys Lys Pro
AAT-CAA-ATG-ATG-AAG-TCT-AGA-AAC-CTC-ACC-AAG-GAC-CGT-TGC-AAG-CCC-
TTA-GTT-TAC-TAC-TTC-AGA-TCT-TTG-GAG-TGG-TTC-CTG-GCA-ACG-TTC-GGG-
                     Xba I                Sau96 I

Val Asn Thr Phe Val His Glu Ser Leu Ala Asp Val Gln Ala Val Cys
GTT-AAC-ACT-TTT-GTG-CAC-GAA-TCC-TTA-GCG-GAT-GTG-CAA-GCC-GTT-TGC-
CAA-TTG-TGA-AAA-CAC-GTG-CTT-AGG-AAT-CGC-CTA-CAC-GTT-CGG-CAA-ACG-
 Hpa I           HgiA I  Hinf I Dde I   Fok I

Ser Gln Lys Asn Val Ala Cys Lys Asn Gly Gln Thr Asn Cys Tyr Gln
AGC-CAA-AAA-AAC-GTT-GCA-TGC-AAG-AAT-GGC-CAA-ACA-AAC-TGT-TAC-CAA-
TCG-GTT-TTT-TTG-CAA-CGT-ACG-TTC-TTA-CCG-GTT-TGT-TTG-ACA-ATG-GTT-
Bbv I                Sph I          Bal I  Tth 111II
Fnu 4HI

Ser Tyr Ser Thr Met Ser Ile Thr Asp Cys Arg Glu Thr Gly Ser Ser
TCG-TAC-TCA-ACT-ATG-TCG-ATC-ACA-GAC-TGC-AGG-GAG-ACT-GGA-AGC-TCA-
AGC-ATG-AGT-TGA-TAC-AGC-TAG-TGT-CTG-ACG-TCC-CTC-TGA-CCT-TCG-AGT-
  Rsa I          Taq I  Mbo I        Pst I               Alu I

Lys Tyr Pro Asn Cys Ala Tyr Lys Thr Thr Gln Ala Asn Lys His Ile
AAA-TAT-CCA-AAC-TGC-GCA-TAT-AAA-ACT-ACC-CAG-GCA-AAC-AAA-CAC-ATC-
TTT-ATA-GGT-TTG-ACG-CGT-ATA-TTT-TGA-TGG-GTC-CGT-TTG-TTT-GTG-TAG-
                 Hha I               BstN I

Ile Val Ala Cys Glu Gly Asn Pro Tyr Val Pro Val His Phe Asp Ala
ATC-GTC-GCG-TGT-GAA-GGT-AAC-CCC-TAT-GTC-CCG-GTT-CAC-TTT-GAC-GCA-
TAG-CAG-CGC-ACA-CTT-CCA-TTG-GGG-ATA-CAG-GGC-CAA-GTG-AAA-CTG-CGT-
        FnuD II     Bst EII             Hpa II       Hga I SfaN I

Ser Val End End    3'
TCT-GTC-TAA-TAA-G      5'
AGA-CAC-ATT-ATT-CCTAG
               BamH I
```

Fig. 5. A sequence of a synthetic gene that codes for RNase. The gene was synthesized chemically, and cloned and expressed in the bacterium E. coli. The underlined regions designate positions where codons were chosen to introduce unique restriction sites into the DNA molecule. These sites permit the chemist to rapidly alter the structure of the gene, and thus synthesize structurally altered proteins.

8. Conclusion

Chemistry has far to go before it exhausts the potential of structural theory. However, to apply structural theory to large molecules, information and tools from other disciplines can be of great value. Synthetic chemistry, molecular biology, and evolutionary theory are providing new insight into the chemistry of proteins, and the processes by which Nature has engineered their properties.

Understanding how Nature altered the behavior of proteins by creating structural alterations is the first step in developing the rules needed by chemists to do the same.

Acknowledgement: The author is indebted to several coworkers, including Dr. Krishnan Nambiar, Dr. Joseph Stackhouse, Dr. Gerard McGeehan, Dr. Geoffrey Chambers, Scott Presnell, Arthur Glasfeld, Rudolf Allemann, Liming Ge, Andrew Ellington, and Dora Stauffer, whose work is discussed. – This work would have been impossible without the support of the Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung, the National Science Foundation (CHE-8213267 and PCM-8111659), the National Institutes of Health (GM-30110), and many private organizations, including Sandoz, Ciba-Geigy, Hoffmann-La Roche, the Searle Foundation, Stuart Pharmaceuticals, Alfred Bader, the Dreyfus Foundation, the Petroleum Research Fund, the Research Corporation, the Clay Science Fund, the Anna Fuller Fund, the Milton Fund, Owens Corning Fiberglass, the Xerox Corporation, the Harvard Society of Fellows, and a gift in memory of Baron Langer von Langensdorf.

Received: March 17, 1987 [FR 37]

[1] a) S. Gillam, K. Waterman, M. Smith *Nucleic Acids Res. 2* (1975) 625; b) G. Dalbadie-McFarland, L.W. Cohen, A.D. Riggs, C. Morin, K. Itakura, J.H. Richards, *Proc. Natl. Acad. Sci. U.S.A. 79* (1982) 6409.

[2] a) R.B. Wallace, M. Schold, M.J. Johnson, P. Dembek, K. Itakura, *Nucleic Acid Res. 9* (1981) 3647; b) S. Inouye, X. Soberon, T. Franceschini, K. Nakamura, K. Itakura, M. Inouye, *Proc. Natl. Acad. Sci. U.S.A. 79* (1982) 3438; c) I.S. Sigal, B.G. Harwood, R. Arentzen, *Proc. Natl. Acad. Sci. U.S.A. 79* (1982) 7159; d) G. Winter, A.R. Fersht, A.J. Wilkinson, M. Zoller, M. Smith, *Nature (London) 299* (1982) 756; e) J.E. Villafranca, E.E. Howell, D.H. Voet, M.S. Strobel, R.C. Ogden, J.N. Abelson, J. Kraut, *Science 222* (1983) 782.

[3] a) M. Kimura, *Nature (London) 217* (1968) 624; b) T.H. Jukes, *Science 210* (1980) 973; c) M. Kimura: *The Neutral Theory of Molecular Evolution*, Cambridge University Press, New York (1983).

[4] a) K. You, *Methods Enzymol. 87* (1982) 101; b) K. You, L.J. Arnold, W.S. Allison, N.O. Kaplan, *Trends Biochem. Sci. 3* (1978) 365.

[5] a) S.A. Benner, *Experientia 38* (1982) 633; b) S.A. Benner, J. Stackhouse, «Chemical Approaches to Understanding Enzymic Catalysis», in B.S. Green, Y. Ashani, D. Chipman (Ed.): *Proc. 26th OHOLO Conference, Zichron Yaacov (Israel)*, Elsevier, Amsterdam (1982), p.32-43; c) K.P. Nambiar, D.M. Stauffer, P.A. Kolodziej, S.A. Benner, *J. Am. Chem. Soc. 105* (1983) 5886.

[6] R.M. Garavito, M.G. Rossmann, P. Argos, W. Eventoff, *Biochemistry 16* (1977) 5065.

[7] J. Stackhouse, K.P. Nambiar, J. Burbaum, D.M. Stauffer, S.A. Benner, *J. Am. Chem. Soc. 107* (1985) 2757.

[8] S.A. Benner, K.P. Nambiar, G.K. Chambers, *J. Am. Chem. Soc. 107* (1985) 5513.

[9] R. Bentley, *Mol. Asymm. Biol. 2* (1970) 1–89.

[10] J.J. Beintema, J.A. Lenstra, in M. Goodman (Ed.): *Macromolecular Sequences in Systematics and Evolutionary Biology*, Plenum, New York (1981).

[11] A. DiDonato, G. D'Alessio, *Biochem. Biophys. Res. Commun. 55* (1973) 919.

[12] S. Vescia, D. Tramontano, G. Augusti-Tocco, G. D'Alessio, *Cancer Res. 40* (1980) 3740.

[13] K. Kurachi, E.W. Davie, D.J. Strydom, J.F. Riordan, B.L. Vallee, *Biochemistry 24* (1985) 5494.

[14] G.J. Gleich, D.A. Loegering, M.P. Bell, J.L. Checkel, S.J. Ackerman, D.J. McKean, *Proc. Natl. Acad. Sci. U.S.A. 83* (1986) 3146.

[15] K.P. Nambiar, J. Stackhouse, D.M. Stauffer, W.G. Kennedy, J.K. Eldredge, S.A. Benner, *Science 222* (1984) 1299.

[16] K. Nagai, H.C. Thogersen, *Nature (London) 309* (1984) 810.