SEQUENCE COMPARISON AND THE PREDICTION OF PROTEIN FUNCTION

by

Gary Edward Otto
B.A., Haverford College
(1978)

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1986

Signature of author_____
Department of Biology
July 1986

Certified by_____
Richard Hynes
Thesis Supervisor

Accepted by_____
David Botstein
Chairman, Biology Department Graduate Committee

SEQUENCE COMPARISON AND THE PREDICTION OF PROTEIN FUNCTION

by

GARY EDWARD OTTO

Submitted to the Department of Biology in August, 1986
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy

ABSTRACT

Two consensus patterns are presented and proven to be predictive of ATP binding. One pattern was wholly derived by comparing the sequences of ATP-binding proteins with structural data describing the nucleotide-binding sites of four well-characterized proteins Pattern elements corresponded with those residues known to contact the bound ligand. While these residues were close to each other in the tertiary structures, they were widely spaced within the sequences. A second pattern was derived in a similar manner by refining a published consensus pattern which was originally identified on the basis of sequence homology. Each pattern was strictly defined by the identities and relative positions of pattern elements. A quantitative measure of the goodness of fit between a pattern and an observed match was derived which is general for any consensus pattern. With this measure, probabilistic predictions of ATP binding could be made which were readily interpretable It was found that 1/3 of all sequenced ATP-binding proteins contained significant matches to one pattern and that an additional 1/3 contained significant matches to the second. Statistical tests of these results demonstrated the correlation between these patterns and the binding of ATP Consensus patterns describing both GTP-binding sites and the NTP-binding sites of DNA polymerases were also derived and these are very similar to one of the patterns for ATP-binding proteins In general, proteins sharing a pattern were not found to share additional homologies by pairwise comparisons. These results are discussed in terms of protein evolution.

The methods of pairwise comparison are discussed at length A method is proposed which allows the significance of an alignment of two sequences to be tested even when that alignment contains gaps. The application of this method to the comparison of groups of sequences and the derivation of consensus patterns is also discussed.

Thesis supervisor: Dr. Richard Hynes
title: Professor of Biology

Acknowledgements

There are many people to thank for discussions of this work and for making my stay at MIT enjoyable. A few who were coerced into longer and more frequent discussions were Jay Sulzberger, Richard Mann, Ihor Lemischka, Jean Schwarzbauer, Chris Kaiser, Parmjit Jat, Jack Price, David Riceman, Jim Thomas, and Suzanna Lewis However, Richard Hynes endured not only the discussions but also the written word His comments and support of this work were very helpful

Table of Contents

List of Figures

List of Tables

INTRODUCTION

Technical advances in molecular biology have made protein sequencing relatively rapid and straightforward As a result, sequences are frequently the first and sometimes the only biochemical data available for proteins of interest. These data can often yield useful information when compared with sequenced proteins of known function. However, there remains the question of how methods of sequence comparison can be made more generally useful in the prediction of protein function. The overall utility of sequence comparison is dependent on two factors 1) the sensitivity of the methods for comparing sequences, and 2) the extent to which proteins have tended to evolve from ancestral precursors as opposed to arising independently.

This thesis proposes a set of methods for the systematic analysis of sequence data. The set ranges from a refinement of existing methods for the comparison of two sequences to the comparison of groups of sequences and the derivation of consensus patterns characteristic of these groups. Each of the methods is sensitive and should be able to detect even distant evolutionary relationships. An analysis of the sequences of ATP-binding proteins, discussed below, demonstrates that large classes of sequences sharing a common function are in fact evolutionarily related As a result, it is expected that these methods will prove very useful in the prediction of protein function While this work focusses on the predictive powers of sequence

comparison, the methods should also allow a quantitative analysis of molecular evolution.

The methods presented here are unrelentingly quantitative but it is hoped that that will not obscure the major points An effort is made throughout to present this material in a way that appeals to intuition. Mathematical formalisms have been kept to a minimum and in that sense some of what is presented must be seen as conjecture. In these instances, the relevant points tend to be demonstrated by experiment rather than proven mathematically. The only mathematical knowledge required is a familiarity with the binomial formula.

Chapter 1 deals with the comparison of pairs of sequences and the identification of homology Each section deals with seperate topics which can be considered independently for the most part. The first section discusses existing procedures used to align two sequences to maximize the correspondence between their residues. The procedures work equally well for any criterion of correspondence chosen. Section 1.2 discusses these criteria of correspondence focussing on which amino acid replacements should be viewed as conservative substitutions Section 1.3 discusses the interpretation of the results from existing methods for searching for homology. In section 1.4 I propose a solution to the problem of assessing the significance of putative homologies between pairs of sequences This section is particularly important and deals with a number of issues, including·

1) a definition of what constitutes an optimal alignment between two sequences
2) a statistical measure of the significance of these optimal alignments which is a normally distributed variable
3) refinements of the methods in Section 1.1 which allow this optimal alignment to be identified
4) a discussion of the underlying assumptions of sequence comparisons which are called collectively the null hypothesis.

and 5) the increased sensitivity which results from the use of these methods

Section 1.5 applies these methods to the comparison of groups of sequences. This application should allow more sensitive tests of evolutionary relationships than simple pairwise comparisons It allows one to ask whether groups of sequences sharing a common function also share evolutionary kinship. This type of analysis can lead to the derivation of consensus sequence patterns that are characteristic of a given function. Methods for dealing with consensus patterns are discussed further in Chapters 3 and 4 Sections 1.4 and 1.5 are relevant to the problem of comparing many sequences simultaneously but this is discussed only briefly. The topic is complicated and a thorough treatment must be deferred.

Chapter 2 applies the methods of Chapter 1 to the prediction of secondary structure. Secondary structure predictors have a form which is similar to the criterion of similarity between

amino acids discussed in Section 1.2. As a result, the methods I develop in Sections 1.4 and 1.5 are immediately applicable to this problem. With these methods, it should be possible to predict a consensus secondary structure for a group of related sequences and quantify the level of confidence with which these predictions are made. Most of this chapter does not have a direct bearing on the remainder of the thesis but the discussion of the modes of protein divergence in Section 2 1 is relevant to Chapters 3 and 4.

Chapter 3 deals with the comparison of sequences sharing a common function which in this case is the binding of ATP. Because it discusses the treatment of consensus patterns it can be viewed as an extension of the methods in Section 1.5. However, this topic was actually the starting point of my work on sequence comparison and this approach can have a different emphasis. The analysis focusses on the comparison of groups of sequences with structural data relevant to the binding of ATP It was found that residues which contact the ligand tend to be well-conserved and that consensus patterns predictive of ATP-binding could be derived from these conserved residues. Two consensus patterns were derived which together identify 70% of all sequenced ATP-binding proteins. These results are discussed in terms of evolutionary descent from ancestral precursors. Chapter 4 extends this approach to other mononucleotide-binding proteins and an additional method for deriving consensus patterns is described.

Taken together, these methods should provide a powerful paradigm for the analysis of sequence data which is outlined in

in the Discussion. Each of these methods can be implemented by relatively straightforward computer programs. Some of these programs have already been written and the remainder can be created by fairly simple modifications of existing programs Each of these programs will be portable and any one of them could be run on a personal computer. Eventually these programs will be incorporated as subroutines in a comprehensive program which will compare all known sequences. This approach will serve two aims: the prediction of protein function and the analysis of the phylogeny of protein evolution.

Chapter I

A Discussion of Homology Comparisons and Some Possible Improvements

The most commonly used method of sequence comparison is the search for homology. While the finding of homology between two sequences says very little about the three-dimensional structures of the proteins, homology implies both a common ancestry and, with less confidence, a common function. This approach has proved useful on numerous occasions by drawing analogies between pairs of proteins not previously known to be similar. Three examples of this are the findings of homology between the oncogene v-*sis* and platelet-derived growth factor (Doolittle *et al* , 1983), between the oncogene v-*erb-B* and the EGF receptor (Downward *et al*., 1984), and between the oncogenes with tyrosine kinase activity and cAMP-dependent protein kinase (Barker and Dayhoff, 1982) These homologies provide strong support for the hypothesis that oncogenic transformation can result from perturbations of the normal signalling mechanisms of the cell.

Homology searches represent an extreme of sequence comparison. With the exception of the sequences themselves, searches are performed with a minimum of additional biological information. Despite the simplicity of the approach, the methodology of homology searches has some complications The major complications involve the evaluation of the statistical

significance of sequence alignments. These complications cause few problems when two sequences are highly homologous However, problems do arise when one tries to identify more distant homologies As a basis for development of new approaches, existing methods of aligning sequences and scoring them for homology will be discussed in Sections 1.1 and 1.2 The interpretation of the results of homology searches is critically discussed in Section 1.3 Sections 1.4 and 1.5 describe my refinements of existing methods aimed at increasing the sensitivity of homology searches.

There are two steps in identifying homologies between sequences, they are: 1) the finding of the best alignment between two sequences, and 2) the evaluation of the statistical significance of the alignment. Most homology algorithms are designed to perform these steps separately. Two types of methods for producing an alignment predominate, they are known as "regions" methods and dynamic programming techniques. Regions methods are the quicker of the two approaches and are commonly used to search large sequence databases for homologs to a sequence of interest. Dynamic programming methods were designed to be more sensitive and to produce optimal alignments, but require much more computing time to perform Both types of methods have been reviewed; the most concise and helpful review is by Waterman (1984), but there are others (Sankoff and Kruskal, 1983; and two issues of Nucleic Acids Research, Vol. 10, No. 1, 1982 and Vol. 12, No. 1, 1984).

1.1 1 The finding of alignments between sequences with regions methods: A popular and representative example of regions methods is the FASTP compucer program by Lipman and Pearson (1985). In part, this program can be characterized as an automated version of dot matrix methods. Dot matrix methods, such as those developed by Maizel and Lenk (1981) and Staden (1982), compare each residue of one sequence along the x-axis of a matrix with the residues of a second sequence along the y-axis Matching residues are identified with dots in the body of the matrix and long diagonals of dots represent homology between the sequences The first step of the FASTP program follows this approach in that the best matching diagonals are identified by numerical methods. To speed the process, only matches between identical residues are scored in this step. Mismatches are allowed but not deletions or insertions. Moreover, the fraction of matching residues in these diagonals must be significantly above the fraction expected for random sequences. A maximum of 5 diagonals are identified in this step. This initial procedure is fast and allows the rapid screening of a database for sequences which may prove to be homologous to the query sequence.

The sequences selected in the first step are analyzed further. The regions of highest matching are rescored to reflect any conservative substitutions between the two sequences. Each corresponding pair of residues in these 5 regions is now scored for similarity using a replaceability matrix (R-matrix). The R-matrix supplies a score for each pair of aligned residues which is meant to reflect their similarity. Scores for identical pairs

of residues or conservative substitutions are positive integers ranging from 1 to 17, the scores of dissimilar pairs are negative integers ranging from -1 to -8. The overall scores for matching regions are the sum of the scores of all residue pairs within the region. While the choice of a scoring method as the basis of comparison is important, the choice does not alter the basic structure of the program and will be discussed in more detail in Section 1.2.

The score of the best diagonal is used as the measure of similarity between the query sequence and sequences in the database, this is known as the initial score. Initial scores for all sequences in the database are stored and their statistical distribution is displayed in a histogram showing the number of sequences with a given initial score This histogram, along with the mean and standard deviation of initial scores, allows the tentative identification of sequences which are homologous with the target sequence. However, the distribution of initial scores is not a normal distribution and this fact complicates the interpretation of results as discussed in Section 1.3.

In the final step of the program, sequences with high initial scores are analyzed further. The sequences are aligned with a procedure which allows for insertions and deletions. The basic correspondence between the sequences is that established by the analysis of long matching diagonals. The final step attempts to identify the longest and best alignment of the two sequences by connecting the best matching diagonals, if possible, using insertions where required. The region of best alignment is scored

again using the R-matrix, penalizing for any insertions used in aligning the sequences. This is the optimal score. For sequences known to share homology, such as the family of oncogenes sharing homology with cAMP-dependent protein kinase, optimal scores are generally higher than initial scores by 30% or more. This improvement in the score after optimization reflects additional correspondences between the sequences beyond the regions of best matching. Along with the initial and optimal scores, the program also displays alignments between sequences. Because of the shortcuts in this algorithm which allow the rapid search of databases, the displayed alignments are not necessarily the optimal alignments of two sequences.

The ability of this program to detect significant homologies is clear. For example, the program identifies the homology shared by protein kinases as well as the homology shared by the family of serine proteases and distinguishes members of these families from the remaining sequences in the database. The initial scores for all of the protein kinase homologs and for the majority of serine proteases are greater than the mean score for the entire database by more than 7 standard deviations. In contrast, the highest scoring sequences which are presumably unrelated to these homologous families lie in the range of 5 to 6 standard deviations from the mean.

The program's limits of sensitivity are more difficult to assess. However, features which distinguish clearly homologous sequences can be tentatively identified by comparing the alignments of the least similar of the related sequences with

those of the most similar of the unrelated sequences. One feature, the length of the matching region, appears to be dominant in distinguishing alignments which clearly reflect homology. The average length of matching regions is 150 residues among related sequences, such as the serine proteases and tyrosine kinases, while it is just 35 among unrelated sequences. In contrast, the quality of matching, as reflected by the percentage of identically matching residues and conservative substitutions, is very similar between the two sets of sequences. There is an average of 27% identical matching and 24% conservative substitutions among related sequences compared with 28% and 21% among unrelated sequences. This issue of the length versus the quality of a matching region and its relation to sensitivity are discussed further in Section 1.4.

As an aside, a more general test of a program's limits of sensitivity could be developed which compares progressively randomized sequences. Beginning with a pair of identical sequences, progressively less homologous pairs could be generated by randomly changing a few residues in turn. These changes could also include the replacement of residues with gaps. Knowing the correct alignment of these sequences beforehand, the quality of matching can be determined for each pair of sequences and the program's ability to find these alignments can be tested. The dependence of sensitivity on the length of the matching region could also be determined in a similar manner.

1 1.2 Dynamic programming methods: Like regions methods, dynamic programming techniques can be thought of as deriving from dot matrix methods. Instead of filling the body of the matrix with dots wnere residues match and then inspecting this display, these programs assign scores for identical matches and conservative substitutions. By adding scores along diagonals according to the algorithm described below, the best matching diagonal can be identified numerically. The basic method was developed by eedleman and Wunsch (1970) for the comparison of DNA sequences. The method is easily modified for the comparison of protein sequences, but is easier to illustrate with DNA sequences. The aim of the method is to produce the optimal alignment between two sequences with respect to a given scoring system. For the purposes of the program, DNA sequences are thought of as being composed of five letters: T, C, A, and G corresponding to the four bases and Δ corresponding to insertions or deletions (indels). Matching residues receive a score of +1, mismatches receive a score of 0, and indels a score of -1.5. Scores for alignments are defined, as in the regions methods, as the sum of the scores for individual pairs of aligned residues and the goal is to find the alignment with the maximum score.

The analysis is done with a device known as the comparison matrix which is similar to the dot matrix. An example of a comparison matrix for two small DNA sequences is shown in Table 1.1. Each cell of the matrix is identified by a pair of integers (i,j). The numbers in each cell of the matrix are determined by the algorithm for finding alignments and proceeds

Table 1.1: Sample comparison matrix for the Needleman-Wunsch algorithm

	j	0	1	2	3	4	5	6	7
i			A	C	T	A	G	T	A
0		0	-1.5	-3.0	-4.5	-6.0	-7.5	-9.0	-10 5
1	T	-1.5	0	-1.5	-2.0	-3.5	-5.0	-5.5	-7 0
2	T	-3.0	-1.5	0	-0.5	-2.0	-3.5	-4.0	-5.5
3	A	-4.5	-2.0	-1.5	0	+0.5	-1.0	-2.5	-3.0
4	G	-6.0	-3.5	-2.0	-1.5	0	+1.5	0	-1.5
5	T	-7.5	-5.0	-3.5	-1.0	-1.5	0	+2 5	+1.0
6	A	-9.0	-5.5	-5.0	-2.5	0	-1.5	0	+3.5
7	C	-10.5	-7.0	-4.5	-4.0	-1.5	0	-1.5	0
8	A	-12.0	-8.5	-7.0	-4.5	-3.0	-1.5	0	-0.5

The best alignment between these two sequences is:

```
A-C-T-A-G-T-A
  | | | | |
T-T-A-G-T-A-C-A
```

in a series of steps. First, initialize the margins of the comparison matrix in three steps:

1) cell (0,0) gets 0
2) proceeding downward from (0,0), each cell gets the value of the cell above it minus 1.5
3) proceeding rightward from (0,0), each cell gets the value of the cell to its left minus 1.5

When comparing sequences, the numerical value of a cell is dependent on the values of the cells preceding it. This step provides these preceding values for the first residues being compared.

Next, calculate scores for each cell in the body of the matrix starting at (1,1) and proceeding rightward row by row These scores are also calculated in steps which determine the dependency of a cell's value on its predecessors:

1) determine the value of a cell as +1 for a match and 0 for a mismatch
2) add the value in (1) to the score of the cell above but subtract 1.5
3) add the value in (1) to the score of the cell to the left but subtract 1.5
4) add the value in (1) to the score of the cell above left (ie. at 350°)
5) choose the highest score out of (2), (3), or (4) and assign it to the cell being scored

Determining the value of a cell by step 4 corresponds to the extension of a matching diagonal. Determining the score of a cell

by either step (2) or (3) corresponds to the insertion of a gap in the alignment, and this insertion is penalized by subtracting 1.5. Note that cells at the end of a long matching diagonal will have high scores. For each cell, the program also records which preceding cell was used to determine its score, producing a path of similarity through the comparison matrix.

Having calculated values for the elements of the comparison matrix, the program surveys them for elements with high scores. In Table 1.1, the highest scoring element is $S_{6,7}$= +3.5. Tracing backwards from this point the program finds a string of five matching residues: TAGTA. This is the optimal alignment. In general, dynamic programming methods can find the alignment between two sequences which is optimal for a given scoring system. These programs can also identify additional regions of potential homology which might be of interest by tracing backwards from other high scoring cells. An example of this is the finding that the sequences of concanavalin A (Con A) and favin are related by a cyclic permutation: the N-terminus of Con A is homologous with the C-terminus of favin while the C-terminus of Con A is homologous with the N-terminus of favin (Erickson and Sellers, 1983).

As was the case with regions methods, the choice of scoring system does not alter the operation of the basic algorithm. Given a scoring system, the algorithm traces the best path through the comparison matrix and this path represents the optimal alignment of the two sequences with respect to the scoring system. Both the FASTP and Needleman-Wunsch programs score on the basis of

similarity and search for alignments with the maximum similarity score. A number of other dynamic programming methods take the opposite approach. They score on the basis of distance, ie. how many mutations are required to convert one sequence into another (Erickson and Sellers, 1983) Optimal alignments are those with minimum distance scores. The intended net effect of dynamic programming methods compared with the regions methods is to allow the identification of alignments with a greater frequency of gaps. Further refinements of this basic algorithm have been made which allow the insertion of larger gaps into sequence alignments (Waterman et al , 1976).

Having found the optimal alignment and calculated its score, the next step is determining the significance of the alignment. This is done by randomizing and re-comparing the sequences a sufficient number of times to obtain a good estimate of the mean and variance of the scores expected for random sequences. With these values, the probability of observing the actual alignment can be estimated.

1.2 Methods of scoring of alignments: While the choice of scoring method does not affect the ways in which the alignment procedures operate, it does affect the utility of the results from homology programs. At issue is which criterion of correspondence is best for comparing pairs of aligned residues and whether that criterion can be improved.

There are four types of criteria of correspondence commonly used in comparing protein sequences. The simplest of these

accepts only identically matching residues in the scoring of an alignment. The remaining criteria make use of additional information about protein sequences. One of these defines the similarity of amino acids according to the genetic code, weighting the similarity of pairs of residues by the minimum number of base changes required to convert one residue into another. Another criterion is defined chemically. For example, anionic residues could be considered similar, as could aromatic or cationic residues The empirically defined criterion is the last type and is derived by comparing very homologous sequences and observing the ways in which they have diverged. These studies clearly show that some amino acids frequently do substitute for some others during evolution. The discussion will focus on empirically defined criteria because they appear to be the most useful.

Two groups have studied the frequencies with which amino acids substitute for each other among highly homologous sequences (Dayhoff et al., 1978; McLachlan, 1971). The best documented of these are the studies by Dayhoff and coworkers which will be the focus of this discussion. They derive a criterion of correspondence from their studies which is the R-matrix referred to above. The steps of this derivation will be outlined and commented upon.

The first step in the process is the tabulation of amino acid substitutions between closely homologous sequences. These data are reproduced in Table 1.2 as the mutation probability matrix. First note that those substitutions which make sense

Table 1 2· Mutation probability matrix for the evolutionary distance of 1 PAM

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	2	1	19	4	1	4	6	1	8	0	1
N Asn	4	.	9822	36	0	4	6	6	21	3	1	.3	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	.	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	5	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

An element of the matrix gives the probability that an amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM. The elements shown are multiplied by 10,000 to simplify the matrix As an example of the use of the matrix, the probability that an Arg residue will be replaced by a Lys is 0 0056 The matrix is reproduced from Dayhoff et al., 1978.

chemically are observed more frequently than others. For example, anionic residues readily substitute for one another as do cationic and aromatic residues. The same observations hold true in the study by McLachlan. Further, Dayhoff et al. note that the observed frequencies of substitution are not what one expects from considering the genetic code and predicting the frequencies of substitutions simply from the minimum number of base changes required for those substitutions. In general, the observed substitutions result from single base changes but not all substitutions arising from single base changes are observed. Therefore, the view of the authors is that the sequences analyzed diverged from one another under conditions of substantial selective pressure and that the observed replacements reflect conservative substitutions of chemically similar amino acids. This view is supported by the agreement between observed substitutions and chemical intuition

However, there are some concerns with these studies. The first is that there are no explicit statistical analyses of the patterns of substitution. Many of the substitutions occur with a frequency much higher than expected from random mutation and they probably do reflect the similarity of some amino acids. That is not clear with all of the substitutions, however, and the apparent similarities and dissimilarities between some amino acids might be the result of random fluctuation. The second issue is the choice of which homologous sequences to compare. There appears to be a bias in the published studies towards using the sequences of extracellular proteins as well as multiple sequences

from single protein families (multiple immunoglobulin sequences, for example). These choices could skew the results. For example, the amino acids Cys and Trp were observed to be strikingly invariant among the sequences compared. Considering the importance of disulfide bridges to extracellular proteins, the invariance of Cys is not surprising. However, it is not clear that this invariance will hold for intracellular proteins.

Another potential bias is the duplication of structural motifs within a protein which are likely to have arisen from a common precursor. This bias is also illustrated by the immunoglobulin light and heavy chain sequences included in the analysis. These proteins are composed of structurally homologous domains, three domains per light chain and four per heavy (Lesk and Chothia, 1982). A central feature of these domains is a disulfide "pin" composed of two Cys residues with a Trp residue in contact with the disulfide bridge. The repeated occurrence of this structural motif in a single class of sequences might be largely responsible for the observed invariance of Trp and Cys residues.

The data in Table 1.2 are the empirical basis for the replaceability matrix (R-matrix). However, these data were put through a series of mathematical manipulations to produce the R-matrix. The mutation probability matrix in Table 1 2 was generated from closely related sequences. The authors calculate the evolutionary distances between the sequences used to generate this matrix as 1 PAM. The PAM unit is defined as percentage of accepted point mutations per 100 residues per 100 million years,

with the term "accepted point mutation" referring to mutations which have become fixed in a population. The PAM-1 matrix (Table 1.2) is meant to embody a statistical model of how proteins mutate and diverge. Divergence of a sequence by one PAM can be simulated using this matrix A random number between 0 and 1 is generated for each residue of the sequence and this number is used to determine the "fate" of that residue in the simulated process of divergence. As an example, consider an alanine residue in the diverging sequence. Referring to the first column of Table 1.2 and proceeding downward from the top, if the random number is between 0 and 0.9867 the Ala remains unchanged, if the number is .9868 the new residue is Arg, if the random number is 0.9869-0.9872 the residue becomes Asn, etc. This procedure is repeated for each residue of the sequence. Lengthier periods of evolution are modeled by successive applications of the matrix to the sequences resulting from the preceding application. For example, the simulation of 10 PAM's worth of divergence would require 10 successive applications of the PAM-1 matrix.

By a similar process, the PAM-1 matrix is converted to matrices which model more distant relationships. The matrix which the authors derive and use is the PAM-250 matrix. This matrix represents a divergence of 250 PAM units and is obtained by multiplying the PAM-1 matrix with itself 250 times. The matrix is shown in Table 1.3. While the choice of the PAM-250 matrix as a basis for comparing sequences, as opposed to a PAM-100 or PAM-200 matrix, is arbitrary there is some justification for it. When proteins have diverged by 250 PAM's only approximately 20% of

Table 1 3· Mutation probability matrix for the evolutionary distance of 250 PAM

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	1	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	22	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	5	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	5	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	6	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

An element of the matrix gives the probability that an amino acid in column j will be replaced by the amino acid in row i after a divergence of 250 PAMs. The elements shown are multiplied by 100 to simplify the matrix. As an example, the probability that an Arg residue will be replaced by a Lys is 0.18. The matrix is reproduced from Dayhoff et al., 1978.

Table 1 4· The replaceability or R-matrix

C	Cys	12																		
S	Ser	0	2																	
T	Thr	-2	1	3																
P	Pro	-3	1	0	6															
A	Ala	-2	1	1	1	2														
G	Gly	-3	1	0	-1	1	5													
N	Asn	-4	1	0	-1	0	0	2												
D	Asp	-5	0	0	-1	0	1	2	4											
E	Glu	-5	0	0	-1	0	0	1	3	4										
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

This is the scoring matrix used in the FASTP program. Positive scores are given to identically matching residues and conservative substitutions. Dissimilar residues have negative scores The matrix is reproduced from Dayhoff et al , 1978.

their residues are expected to be unchanged (Dayhoff et al., 1978). It is at these evolutionary distances that one most needs an additional basis of comparison beyond that of simple identical matching. Note that while divergence by 250 PAM's means that a span of 100 residues has undergone 250 mutations, 20% of the residues remain unchanged. This is because an increasing number of these mutations occur at sites previously mutated as the process of divergence proceeds.

The derivation of the PAM-250 matrix causes some concern because it is based on the unproven assumption that mutation is, on average, a constant process. This assumption could be tested. Collect the alignments of pairs of sequences which are clearly homologous by the criterion of identically matching residues. Rank these alignments in order of increasing divergence as measured in PAM units. Separate this ranked list into groups of alignments with similar PAM values and generate a mutation probability matrix for each group. Each of these groups is characterized by its average PAM value and the question is whether the mutation probability matrices generated by the above procedure are similar to those with similar PAM values generated by matrix multiplication. This approach provides a good test of the underlying assumption of the matrix multiplication method and moreover is itself a good way to define the replaceability of amino acids for one another.

The PAM-250 matrix is the penultimate step in the development of this empirically defined scoring system. The final step is the creation of the relatedness odds matrix; this is the

R-matrix and is shown in Table 1.4. Each term in the R-matrix is the logarithm of the ratio of the observed frequency of substitution between amino acid pairs over the frequency of substitution expected by chance. Substitutions which occur more frequently than expected take on positive integer values, those which occur as frequently as expected take on the value zero, and those occurring less frequently take on negative integer values. Alignments are scored by adding values from the R-matrix for each pair of aligned residues. Thus those residue pairs with positive values support a hypothesis that two sequences have diverged from a common ancestor, those with negative values discount it and those with zero values have no effect.

The next question is how to test this scoring system. The performance of this scoring system was compared with that of other systems (Dayhoff et al., 1978b). These systems included: 1) the scoring of only identically matching residues with all matches receiving a score of 1; 2) a system defined by the genetic code; and 3) a number of empirically derived scoring systems. The R-matrix was found to be the best in these comparative studies. However, it appears that some of the sequences used in the test were also used in the derivation of the R-matrix. Moreover, this procedure does not directly test the central feature of the R-matrix. The primary motivation for creating the R-matrix and not simply scoring identical matches is to use information about conservative substitutions in the scoring of alignments. Therefore a direct test of the off-diagonal matrix elements is required. This could be accomplished

by comparing the off-diagonal scores of relevantly aligned homologous sequences with the scores of randomly generated sequences. The first set of sequences to collect are those which are clearly homologous by the criterion of identical matching and which have not been used in the derivation of the scoring method. The question is then asked whether the scoring of the non-matching residues using the off-diagonal elements of the R-matrix helps or hinders the finding of significant homology. Next, mock alignments of randomly generated sequences are scored in the same manner. I have made some comparisons of this type which are shown in Table 1.5. These results show that the overall effect of these off-diagonal elements is to aid in the identification of homology by causing a dramatic reduction in the scores of irrelevant alignments.

This analysis of replaceability should be repeated using larger sets of homologous sequences of different classes, such as intra- and extracellular proteins. Each class of sequences should be analyzed separately with methods that test the significance of similarities observed between amino acids. By comparing the results obtained with each class, one can test whether there is one R-matrix which is generally applicable to all protein sequences. However, the question is not whether empirically defined scoring systems are valid but whether they can be improved.

The next issue is what a similarity score obtained with the R-matrix actually measures. It is the case that an alignment with a higher score is more likely to reflect a genuine evolutionary

Table 1.5 The contribution of off-diagonal to the overall alignment score obtained by using the R-matrix.

a. Results with members of the homologous family of protein kinases

Sequence	overall score	diagonal[a] elements	off-diagonal[b] elements
v-fes	87	92	-5
v-mht	86	69	17
v-mos	76	71	5
v-fps	93	102	-9
Averages	86	82	4

b. Results with randomly aligned sequences

Sequence	overall score	diagonal[a] elements	off-diagonal[b] elements
1	-29	4	-33
2	-17	12	-29
3	-50	6	-56
4	10	18	-8
5	-20	15	-35
Averages	-21	11	-32

a) sum of the scores of identically matching residues

b) sum of the scores of non-identical pairs including conservative substitutions and mismatches

relationship between two sequences than an alignment with a lower score. However, the score itself is not sufficient to indicate a relationship between two sequences. The relevant question is how much the observed score deviates from the distribution of scores expected for unrelated sequences.

1.3 Interpretation of homology searches: The interpretation of sequence comparisons requires a statistical analysis of the data resulting from a homology program. This section focuses on the analysis of comparisons of single sequences with the entire protein sequence database. Such comparisons are done most commonly with regions methods. Results obtained with the FASTP program will be considered as an example. The FASTP program displays the distribution of similarity scores observed in the comparison of one sequence with all sequences in a database. The mean score and standard deviation are also calculated. However, these data are not generally sufficient to assess the significance of an alignment, additional information is required. This can be illustrated by analyzing the comparison data for trypsinogen and for cAMP-dependent protein kinase (cAPK) presented by Lipman and Pearson (1985).

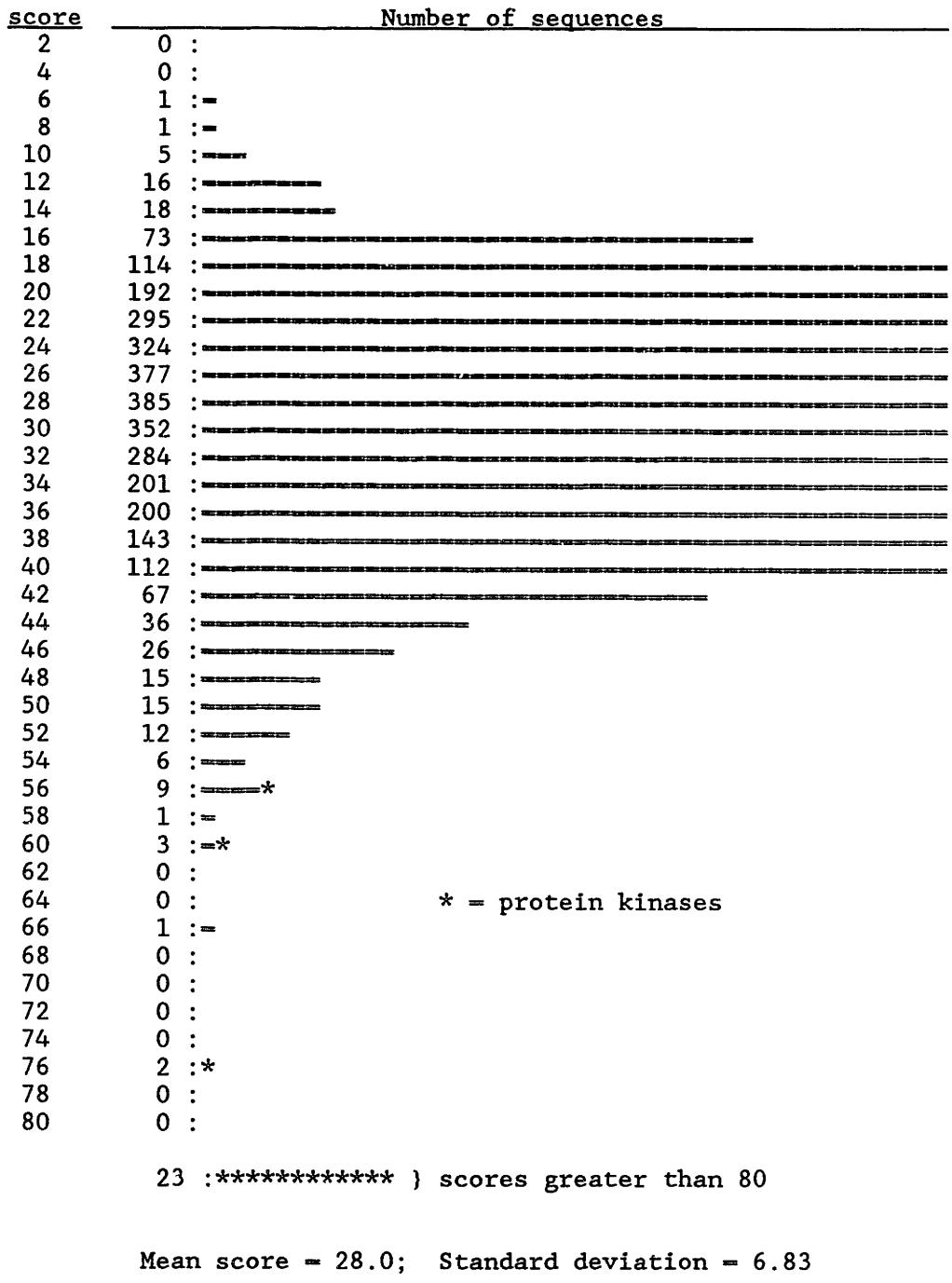
The distribution of similarity scores displayed by the FASTP program for searches of the database with the sequences of trypsinogen and cAPK are shown in Tables 1.6 and 1.7. Sequences known to be related to the query sequences are denoted by an asterisks in these tables. Ignoring these related sequences, the distributions of scores for proteins not known to be related to

Table 1.6 Histogram of homology scores resulting from the comparison of bovine trypsinogen with the sequence database using the FASTP program

score	Number of sequences
2	3 :===
4	1 :-
6	1 :-
8	3 :===
10	51 :==========
12	117 :==========
14	230 :==========
16	423 :==========
18	408 :==========
20	523 :==========
22	412 :==========
24	330 :==========
26	213 :==========
28	235 :==========
30	152 :==========
32	58 :==========
34	48 :==========
36	26 :==========
38	17 :==========
40	5 :====
42	8 :====*
44	3 :===
46	0 :
48	4 :==
50	0 :
52	0 :
54	1 :-
56	1 :-
58	0 :
60	0 :
62	0 :
64	2 :* * = serine proteases
66	0 :
68	0 :
70	1 :*
72	0 :
74	0 :
76	0 :
78	1 :*
80	0 :
	32 :********** } scores greater than 80

Mean score = 20.7; Standard deviation = 5.69

Table 1.7 Histogram of homology scores resulting from the comparison of cAMP-dependent protein kinase with the sequence database using the FASTP program



trypsinogen or cAPK are compared in Figure 1.1. Only the tails of these distributions are displayed since these are the regions relevant to the identification of homology. The large difference between these distributions makes it clear that the similarity score itself is a poor measure of similarity since these scores are not comparable from one comparison to the next.

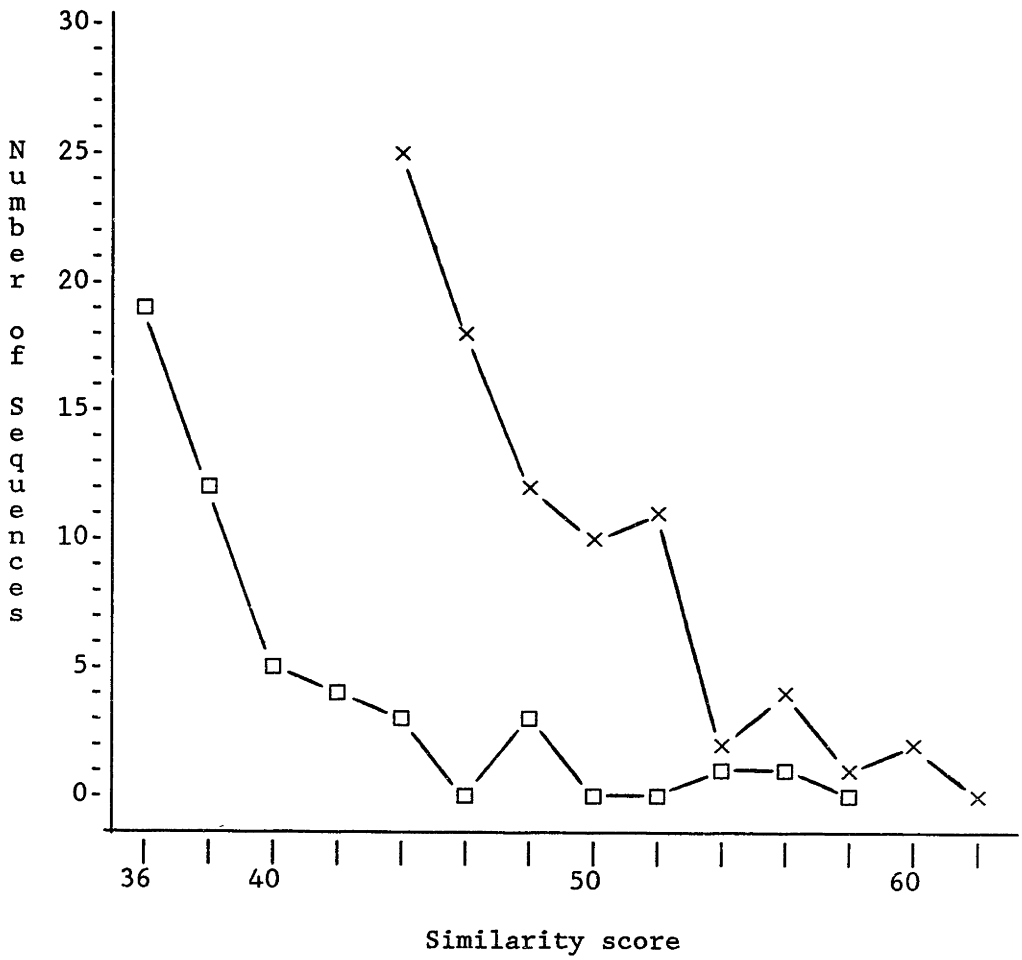
However, these data can be re-analyzed to produce a display of similarity which is more interpretable. The mean and standard deviation reported for each comparison by the FASTP program can be used to convert similarity scores into units of standard deviation (Lipman & Pearson, 1985). When the distributions are re-plotted in these units, as shown in Figure 1.2, the results of the comparison of cAPK and trypsinogen are seen to be very similar. Furthermore, these distributions are very similar to that of a randomly generated sequence, as shown. The conclusion drawn from Fig. 1.2 is that the high scoring alignments in these comparisons do not represent previously undetected homology but are simply the tail of the distribution of scores expected by chance. In contrast, the homologies among protein kinases and serine proteases are emphasized by this type of analysis. When the scores of those alignments are expressed in units of standard deviation they range from 7 to 59.

Ideally, the results of the FASTP program would be displayed as in Fig. 1.2, reporting both the actual value of the similarity score and its value in units of standard deviation for each sequence. The results of searches displayed in this way should be directly comparable. By comparing a number of randomly generated

Figure 1.1: Two distributions of similarity scores resulting from searches of the database with the sequences of cAMP-dependent protein kinase and trypsinogen.

The data from Tables 1.6 and 1.7 are compared in this graph. The scores of serine proteases and protein kinases were excluded. The remaining sequences are not known to be homologous with the query sequences and their scores should reflect the frequency of random matching. Only the tails of their score distributions are shown but these reflect the differences between the distributions as a whole. For example, only 2 sequences from the trypsinogen search have scores of 50 or more compared with 30 sequences from the protein kinase search.

Figure 1.1

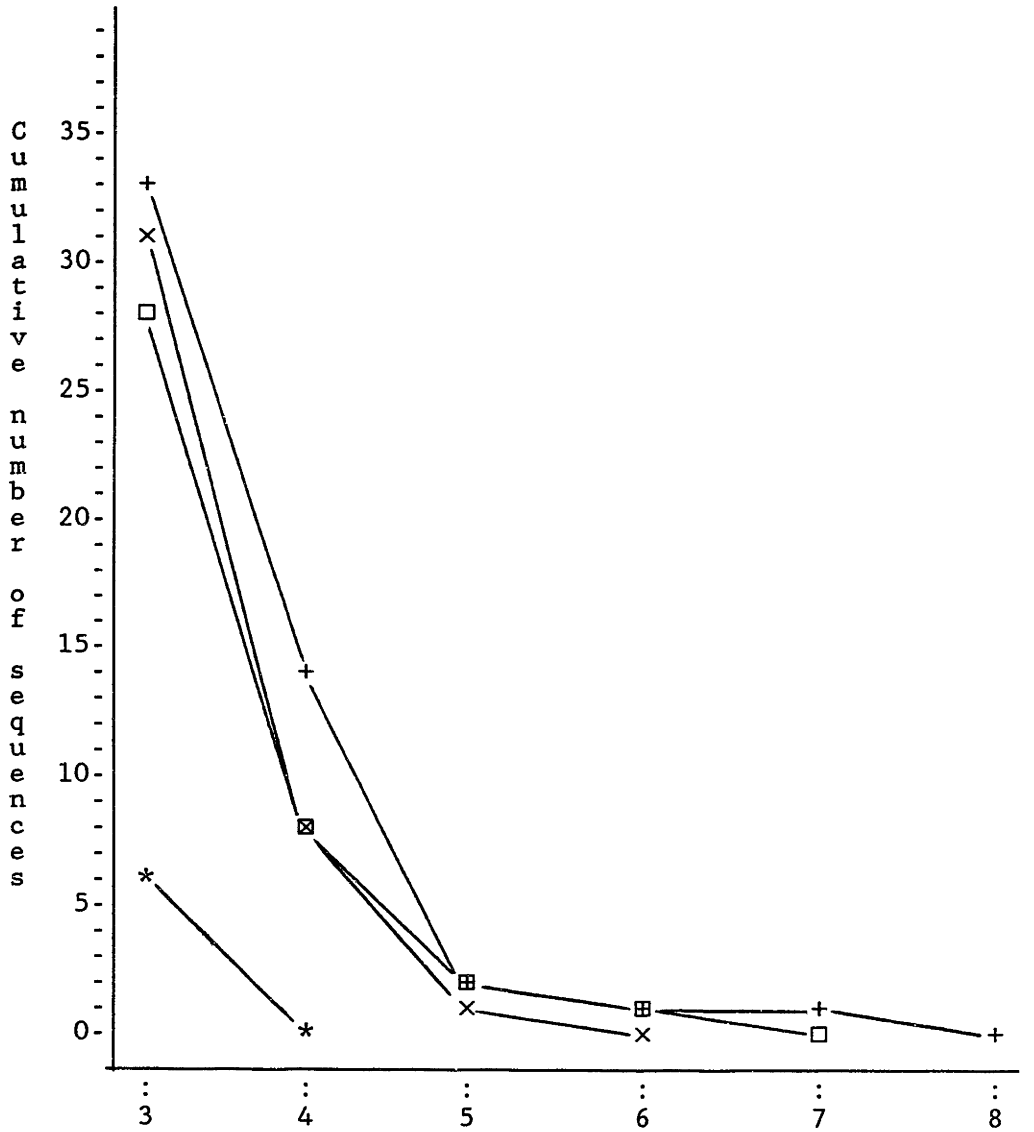


× --- × = cAMP-dependent protein kinase

□ --- □ = trypsinogen

Figure 1.2: Distribution of scores of unrelated sequences expressed in units of standard deviation.

The data in Fig. 1.1 are re-plotted. Here scores are expressed in units of standard deviation, (score - mean)/(standard deviation), and plotted against the cumulative number of sequences. In contrast with Fig. 1.1, the distributions resulting from searches with cAMP-dependent protein kinase and trypsinogen are similar both to themselves and to the distribution from a search of the database with a randomly generated sequence.



× --- × = cAMP-dependent protein kinase

□ --- □ = trypsinogen

\+ --- + = randomly generated sequence

* --- * = results expected with a normally distributed measure of homology

sequences with the database, one can establish the extent of matching expected for unrelated sequences. The results of actual sequence comparisons can then be compared against this standard.

Lipman and Pearson note that the scores resulting from comparisons with the FASTP program are not normally distributed. The distribution of scores expected of a measure of homology which is normal is also shown in Fig. 1.2. With a normal measure of homology, scores 4 or more standard deviations from the mean are not expected among unrelated sequences given the size of the current database of approximately 2500 different sequences. In contrast, scores for unrelated sequences are observed from the FASTP program which are 4, 5, 6, and 7 standard deviations higher than the mean. As a result, a normally distributed variable is likely to be a more sensitive measure of homology than a non-normal measure.

1.4 A normally distributed measure of similarity: The fundamental measure of homology is the probability of an alignment between two sequences occurring by chance. Therefore, it follows that an alignment is optimal only if its associated probability value is the lowest observed in the comparison. The values calculated for the probability of an alignment occurring by chance are dependent on one's assumptions of how proteins are built. The major assumption made here is the statistical independence of each residue in a sequence. A more thorough discussion of these assumptions is deferred to Section 1.4.1.

In the programs discussed above, the alignments sought are those with the highest similarity score. However, an alignment

with the highest similarity score is not necessarily the optimal alignment by probabilistic criteria. Therefore, a refinement of the algorithms described above would search for alignments satisfying the criterion of least probable score. There are established statistical techniques which can be used to evaluate probability values for alignments which do not contain gaps. These methods can be applied to any scoring method such as the replaceability matrix of the FASTP program. I describe a method for assessing the significance of alignments containing gaps. This method is then used to define a measure of similarity which is shown to be normally distributed.

Both regions and dynamic programming methods evaluate alignments as the simple sum of scores for each pair of aligned residues. Evaluating in this fashion, two alignments with matching regions of different length could have the same similarity score and these alignments would be treated as equivalently significant observations by existing methods. However, the probabilities associated with two alignments having the same similarity score are not generally equal, since these probability values are dependent on both the quality and length of the matching region.

While programs which search for probabilistically optimal alignments have yet to be developed, some essential features can be outlined. First, if one assumes that insertions and deletions are relatively rare events when compared with the frequency of substitutions, then the basic algorithm of the FASTP program can simply be modified. Long matching diagonals are searched for just as before except that these diagonals are identified by the

criterion of P(score) rather than the scores themselves. The best matching diagonals are then connected, with the insertion of gaps where required, to produce the optimal alignment between two sequences. A second value, P(align) or P(a), is required to describe the probability of observing an overall alignment which includes gaps.

A method for calculating the probability of any alignment which contains no gaps is given by McLachlan & Boswell (1985). The method is general for any scoring system such as that of the FASTP program. Without going into detail, the method relies on mathematical devices known as generating functions to construct families of functions describing the statistical distribution of scores. Each function in a family is the discrete probability distribution of scores expected for a matching region of a particular size and is dependent on the molar compositions of the two sequences being compared. With these functions, P(score) can be explicitly calculated for a matching diagonal of any size and score. Yet because these functions are dependent on the molar composition of the two sequences being compared, they must be generated again for each new pair of sequences examined. However, efficient methods of approximating P(score) are also available McLachlan & Boswell (1985).

The authors apply this statistical method to the output of a program which assesses homology by comparing sequence segments which are of fixed length. Segments are scored for similarity by summing the contributions of the individual residue pairs. The segment length is an adjustable parameter so that one may chose to compare, for example, all segments of length 5 or all segments

of length 10. The sensitivity of the program can be controlled with this length parameter: the smaller the length, the greater the sensitivity. Naturally, the occurrence of spurious similarities increases as the sensitivity increases. This program is a descendent of the more visual dot matrix program of Staden (1982). With the Staden program, a comparison performed with high sensitivity produces a matrix which is full of dots which tend to obscure any long diagonals which might reflect homology.

Sequences need not be compared by segments of fixed length, however. An alternate approach would search along the diagonals of a comparison matrix and identify segments of any length for which P(score) is a local minimum. In this approach, the adjustable parameter affecting sensitivity is a cutoff value for P(score). This procedure could be readily incorporated into a dot matrix program and the diagonals appearing in these matrices would then represent the most significant similarities between two sequences. An important point to be made about the statistical method of McLachlan & Boswell is that it is essentially exact given the assumptions discussed in the next section. As a result, the choice of a cutoff value for P(score) predicts the number of matching diagonals expected for unrelated sequences. For example, if two sequences of length 100 are compared with a cutoff of P(score)= 0.001, one expects to find approximately 10 matching diagonals (100x100x0.001). How much of an improvement this approach offers in practice over the old one remains to be seen. However, the data displayed with the new approach are clearly the relevant data for assessing homology.

The above methods would be sufficient to assess any pair of sequences for homology if insertions/deletions (indels) never occurred in alignments of homologous sequences. Indels do occur, however, and complicate this assessment. If the number of indels in a valid alignment of homologous sequences were large, the matching diagonals reflecting this homology would be short and none of the values of P(score) associated with these diagonals would, in themselves, be significant enough to indicate homology. In this case, a dot representation of the comparison matrix would show that the relevant diagonals are obscured by the spurious diagonals expected for even random sequences. The question then is how these relevant diagonals can be connected by the insertion of gaps to produce an overall alignment of the two sequences. The programs discussed in Section 1.1 will take each matching diagonal and attempt to connect it with a neighboring diagonal by the progressive insertion of gaps up to some predetermined limit. In the FASTP program, for example, the largest number of gaps which can be inserted is 7. However, this limit need not be predetermined but could instead be a function of the quality of matching observed in the diagonals to be connected. What is required is some concept of "local significance". Each matching diagonal would have adjacent regions of the comparison matrix associated with it. The sizes of these regions would be dependent on the diagonal's value of P(score). The program would then search for neighboring diagonals only within these associated regions. In this way, a set of matching diagonals representing an overall alignment could be selected and tested for significance.

This last step poses a problem because there is not an accepted method for determining the significance of alignments which contain gaps. However, as stated above, probability values can be calculated for any alignment which does not contain gaps. If the gaps in an alignment are simply ignored, one is left with one long matching diagonal the P(score) of which can be calculated. This value, which will be designated by P(s), does not describe the probability of observing the alignment because the gaps have been ignored. However, P(s) is a lower limit for the probability of the observed alignment. The actual probability of an alignment which contains gaps, designated P(a), will always be greater than P(s), ie. it will always be less significant.

The next question is how to calculate P(a) from the value of P(s) and the number of gaps which have been ignored. The first step is to count the number of alternative alignments, with the same or better values for P(s), which can be produced by the manipulation of the gaps. This number will be called G. Consider an alignment which contains some gaps:

```
AQWSRFDY-AKSMGPIH--AIFTDRGWLPGTHRTESDGYNCPQGGPIYTREWQAS-FGHKLM
|||  ||| || || ||  | || | |||   | |  ||  || | | | || || | | ||
AQWGTFDYNAKTMGNIHLQAYFTHRPWLP---RFECSGYMLPQEG-IQTSEWCASFFNHPLM
```

The positions of the gaps are determined by the matching diagonals since repositioning of the gaps would degrade the quality of the alignment and give a poorer value of P(s). Therefore, repositioning of gaps is not allowed in the counting process. Similarly, filling in gaps (for example, DY-AK => DYXAK) is not allowed because the value for P(score) of the resulting long matching diagonal would be greater than P(s). Only the

removal of gaps, along with their corresponding residues in the opposite sequence, is allowed as shown below. Because an alignment would have the same value for P(s) whether a particular gap is in one of the sequences or the other, each arrangement of gaps is counted twice. For a gap two residues long, the number of arrangements is 5:

```
   IH--AIF          IH-AIF
1) ||  | |      2) || | |
   IHXXAYF          IHXAYF          IHAIF
                                 5) ||| |
   IHXXAIF          IHXAIF          IHAYF
3) ||  | |      4) || | |
   IH--AYF          IH-AYF
```

By similar reasoning, the number of possible arrangements per gap is $(2x + 1)$, where x is the number of residue positions per gap. Every valid manipulation of every gap produces an alignment with the same value of P(s). Therefore, the value of G must be a count of all the allowable permutations of the gaps and is calculated as the product of values of $(2x + 1)$ for all gaps. For example, in the above alignment with 5 gaps, the value of G is: $\{(2\times1+1)^3(2\times2+1)(2\times3+1)\} = 945$. Having counted the number of equivalent alternate alignments engendered by the presence of gaps, P(a) can be calculated as a one-tailed binomial probability of finding one or more alignments as good or better as that observed:

$$P(a) = 1 - \{1 - P(s)\}^G$$

Because the value of G increases exponentially with the number of gaps, even a relatively modest number of gaps will greatly reduce the significance of alignments.

While P(a) can be used to identify the optimal alignment between two sequences, this value is still not an adequate measure of homology. This is because values for P(a) describe the probability of observing the alignment in just one trial. However, each diagonal of the comparison matrix represents another alignment which has been examined for similarity: each diagonal is an additional trial. Therefore, a third probability value is needed which describes the likelihood that the best alignment between two sequences represents a significant similarity between them. This value, P(sim), will be dependent on an estimate of the number of trials, T. How P(sim) is best calculated is not immediately obvious. However, a reasonable first approximation might be:

$$P(sim) = 1 - \{1 - P(a)\}^{T}$$

$$T = NM - [N + M - (L-1)](L-1)$$

The first equation is a test of the significance of P(a) given an estimate of the number of trials, T. The second equation is an estimate of T where M and N are the lengths of the two sequences being compared and L is the length of the matching region. T represents the area of the entire comparison matrix minus the

area of the edges of the matrix in which the matching region cannot possibly fit.

The motivation for defining the variable P(sim) as the measure of homology is to simplify the interpretation of searches of a database. P(sim) represents the probability of finding an alignment between randomized sequences which is better than that observed in the actual sequences. If two sequences are unrelated, then a comparison of them will be indistinguishable from a comparison in which these sequences have been randomized. As a result, the best alignment out of either comparison has an equal probability of being found in either one or the other comparison matrix. Therefore, with unrelated sequences, the mean value for P(sim) expected by theory is 0.5. Moreover, values of P(sim) are expected to be normally distributed about this mean of 0.5 because values of P(sim) are the result of a large number of small and independent contributions of individual pairs of aligned residues. If the values of P(sim) for unrelated sequences are normally distributed, the interpretation of homology searches becomes very straightforward, as discussed in Section 1.5.

The above method of analysis was tested with the results of a homology search which compared the sequence of cAMP-dependent protein kinase with the 1985 protein sequence database. The search was done with the FASTP program. After deleting known homologs to cAPK, the remaining sequences produced in the search were assumed to be unrelated to cAPK and 26 of these were used for the test. Values for P(s) based on the program's own scoring criterion, ie. the replaceability matrix, could not be calculated

because the needed method has not yet been implemented. However, values of P(s) based on a criterion of identically matching residues could be calculated. These calculations are entirely analogous with the methods discussed above. In this case, the generating function, which would be very complicated for the replaceability matrix, reduces to the much simpler binomial formula. The value for the frequency of identical matching expected of unrelated sequences used in the following calculations was 0.058. This value is obtained by summing the probabilities of each possible type of match, ie. matching leucines or matching tryptophans. The calculation is analogous to that used by Smith et al. (1985) for DNA sequences and assumes that all proteins share the average molar composition of proteins, values of which are from Dayhoff et al. (1978). The deviation of this value from 0.05 is due to the fact that amino acids do not occur in protein sequences with the same frequency

Two examples of the calculation of values for P(s), P(a), and P(sim) are shown in Table 1.8. The sequences chosen for this example were v-fms and a fragment of human c-abl which were the worst scoring of the tyrosine kinases with similarity scores of 55 and 60, respectively. The values calculated were 2.8×10^{-4} for v-fms and 4.3×10^{-4} for c-abl. Both values of P(sim) are significantly low and indicate homology between these sequences and cAMP-dependent protein kinase. The number of random matches expected between protein kinase and unrelated sequences is obtained by multiplying the above values of P(sim) by the number of unique sequences in the database (2200). For v-fms this

Table 1.8: Sample calculations of values for P(sim)

a. Comparison of cAMP-dependent protein kinase with the tyrosine kinase v-fms

```
cAMP-dep. kinase     QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVK
                     :      : :   : ::    : :          ::: :
v-fms Tyr kinase     QVAQGMAFLASKNCIHRDVAARNVLLTSGRVAKIGDFGLARDIM

GRTWTLCGTPEYL-----APEIILSKGYNKAVDWWALGVLIYEM-AAGYPPFFADQPIQIYEKIV
            :     ::: :    :    : : : :  :    : :          : :
NDSNYIVKGNARLPVKWMAPESIFDCVYTVQSDVWSYGILLWEIFSLGLNPYPGILVNSKFYKLV
```

N = 1511 residues = size of v-fms tyrosine kinase
M = 350 residues = size of cAMP-dependent kinase
L = 103 residues = size of region of alignment
h = 27 residues = identically matching residues

A value for P(s) is first calculated with the binomial formula, the gaps in the alignment are simply ignored. The value used for the frequency of identically matching residues expected of unrelated sequences is 0.058, as described in the text.

$$P(s) = \sum_{i=h}^{L} (.058)^i (0.942)^{(L-i)} \binom{L}{i}$$

$$= \sum_{i=27}^{103} (.058)^i (0.942)^{(103-i)} \binom{103}{i}$$

$$= 2.53\times10^{-11}$$

P(a) takes account of the gaps previously ignored. The alignment contains 2 gaps of size 1 and 5, so G = (2×1 + 1)(2×5 + 1) = 33 and

$$P(a) = 1 - [1 - P(s)]^C$$
$$= 1 - [1 - (2.53\times10^{-11})]^{(33)}$$
$$= 8.33\times10^{-10}$$

Given the sizes of the sequences and the region of alignment,

$$T = NM - [N + M - (L-1)]\times(L-1)$$
$$= (1511)(350) - [1511 + 350 - 102]\times(102)$$
$$= 349,432$$

The value for P(sim) is then calculated from P(a) and T:

$$P(sim) = 1 - [1 - P(a)]^T$$
$$= 1 - [1 - 3.3\times10^{-9}]^{(349,432)}$$
$$= 2.8\times10^{-4}$$

Table 1.8: (continued)
b. Comparison of cAMP-dependent protein kinase with a fragment of the tyrosine kinase c-abl

cAMP-dependent kinase
LKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEY----LAPEIILSKGYNKAVDWWALG
:: : : : ::: : : : ::: : : :
DLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWGKG
v-abl kinase, fragment

N = 64 residues = size of v-abl kinase fragment
M = 350 residues = size of cAMP-dependent kinase
L = 60 residues = size of region of alignment
h = 17 residues = identically matching residues

A value for P(s) is first calculated with the binomial formula; the gaps in the alignment are simply ignored. The value used for the frequency of identically matching residues expected of unrelated sequences is 0.058, as described in the text.

$$P(s) = \sum_{i=h}^{L} (.058)^{i}(0.942)^{(L-i)} \binom{L}{i}$$

$$= \sum_{i=17}^{60} (.058)^{i}(0.942)^{(60-i)} \binom{60}{i}$$

$$= 3.30\times10^{-8}$$

P(a) takes account of the gaps previously ignored. The alignment contains 1 gap of size 4, so G = (2×4 + 1) = 9 and

$$P(a) = 1 - [1 - P(s)]^{G}$$
$$= 1 - [1 - (3.33\times10^{-8})]^{(9)}$$
$$= 2.97\times10^{-7}$$

Given the sizes of the sequences and the region of alignment,

$$T = NM - [N + M - (L-1)]\times(L-1)$$
$$= (64)(350) - [64 + 350 - 59]\times(59)$$
$$= 1,455$$

The value for P(sim) is then calculated from P(a) and T:

$$P(sim) = 1 - [1 - P(a)]^{T}$$
$$= 1 - [\,1 - 2.97\times10^{-7}]^{(1,455)}$$
$$= 4.3\times10^{-4}$$

expected number of random matches is 0.6 and for c-abl it is 1.0. Table 1.7 shows that with the FASTP program, the number of random matches which scored as well or better than the related sequences was 12 for v-fms and 2 for c-abl. This demonstrates that the sensitivity of the comparisons made with the FASTP program is improved by this alternative method of analysis. Moreover, it is important to note that the calculations in Table 1.8 are based on the criterion of identical matching. Extending the method to include conservative substitutions as matches will further increase the sensitivity.

The results of the test on the 26 unrelated sequences are shown in Table 1.9. The overall observed mean is 0.50 and the standard deviation is 0.32; the values expected for a normally distributed variable are 0.5000 and 0.3174, respectively. The striking agreement between the observed and expected values are due to the size of the sample. Assessing the significance of alignments which contain gaps has been an especially intractable problem in the field so a test of this particular aspect of the method was done by analyzing alignments with and without gaps separately. Again, the observed values for mean and variance are in good agreement with those expected by theory.

The effort was made to make the test's sample set unbiased but this involved some selection of sequences. All top-ranking unrelated sequences were selected for the sample set until 16 alignments which contained no gaps were collected, then only alignments with gaps were selected until their number was 18. Homologous sequences, such as a group of hemoglobins, were

Table 1.9 Assessment of the probabilities of alignments between cAMP-dependent protein kinase and sequences from the sequence database occurring by chance

Sequence	% identity	size of overlap region	P(s)	G	P(a)	T	P(sim)
Hypoth. B256	25.5	47	1.1e-5	1	1 1e-5	63327	0.51
β-Tubulin	39.1	23	2.9e-6	1	2.9e-6	141918	0.33
α-Crystallin	26.2	61	2.7e-7	15	4.0e-6	31501	0.12
Amidase	31.0	29	2.6e-5	1	2.6e-5	39162	0.63
EpBarr BNLF	36.0	25	6.4e-6	1	6.4e-6	117325	0.53
Hemoglobin	31.4	35	2.8e-6	3	8.5e-6	34965	0.26
Anthran. syn.	26.9	52	1.0e-6	3	3.2e-6	139464	0.36
Gene 0.6	35.0	20	8.7e-5	1	8.7e-5	30030	0.93
Egg	38.1	21	1.3e-5	1	1.3e-5	110544	0.76
Complem. C4a	35.7	28	2.1e-6	5	1.1e-5	545468	0.997
Ea22 gene	36.0	25	6.5e-6	1	6.5e-6	51025	0.28
Hypoth. C130	45.0	20	6.9e-7	1	6.9e-7	36300	0.025
Bo-LV polym.	38.1	21	1.3e-5	1	1.3e-5	273399	0.97
oxi 3	19.8	162	1.1e-9	15309	1.7e-5	40608	0.53
Proenkephalin	34.4	32	1.0e-6	1	1.0e-6	74730	0.072
AL1	42.9	21	1.2e-6	1	1.2e-6	108899	0.12
Fumarate red.	31.2	32	8.5e-6	1	8.5e-6	67416	0.43
Complem. C5a	31.8	22	1.7e-4	1	1.7e-4	17056	0.95
tail Z	30.3	33	1.2e-5	3	3.5e-5	50403	0.82
Cytochr. oxid.	24.1	79	9.6e-8	27	2.6e-6	40379	0.10
Enolase	21.2	99	2.1e-7	55	1.1e-5	84587	0.62
Tet. resist.	26.2	61	2.7e-7	21	5.6e-6	97682	0.42
L pro	24.5	98	1.4e-9	105	1.5e-7	506772	0.07
A mos V	43.8	16	1.6e-5	1	1.6e-5	258516	0.98
Tail K	20.0	120	9.6e-8	729	7.0e-5	18170	0.72
dnaQ	21.5	102	8.0e-8	315	2.5e-5	34968	0.58

	Mean	Standard deviation
Alignments with no gaps (14)	0.54	0.34
Alignments containing gaps (12)	0.47	0.30
All alignments (26)	0.50	0.32

ignored. Eight of the 34 collected sequences had P(sim) values of precisely 1, given the six decimal precision of these calculations, compared with 0.997 for the highest valued alignment in Table 1.9. Therefore, these 8 alignments are viewed as outliers, ie. atypical of the distribution shared by the majority of the alignments, and were excluded from the statistical analysis. The deviation of P(sim) for these 8 sequences from that expected is accounted for by the disproportionality of the FASTP scoring system. Each of these alignments contain a high proportion of Phe, Tyr, and Trp residues which are given very high similarity scores in the R-matrix and are mainly responsible for the program's selection of these sequences.

1.4.1 Assumptions underlying the comparison of sequences: The assumptions underlying the statistical analysis of homology have been largely ignored in the preceding discussion. There are two major assumptions which will be discussed. The first is that every protein can be considered as a sequence of residues drawn independently from a pool of residues, "with replacement". Second, this pool of residues is assumed to be identical for every protein and to share the average molar composition of all proteins. The results in Section 1.4 show that these assumptions are reasonable to a first approximation. However, the results of future research will cause these assumptions to be modified, as discussed below.

When assessing the significance of a sequence alignment, the value of interest is the probability of finding an alignment between random sequences as good or better than that observed between the actual sequences. This value can be obtained by comparing many pairs of random sequences which have been constructed by following a set of rules. There are two different types of sets of rules. The most commonly used set of rules constructs random sequences by shuffling the residues of the actual sequences. In this instance, residues are treated as statistically independent objects but the sampling is from a finite pool, ie. from the actual sequences, and is done without replacement. The exact compositions of the two sequences are maintained in this process. An alternative set of rules constructs random sequences by sampling from an infinite pool of residues which reflects the molar composition of the two sequences being compared. In effect, it is sampling with replacement. This second method of constructing random sequences is the easiest to model with probability theory. It is under this assumption that the generating functions presented by McLachlan & Boswell (1985) and the simpler binomial calculations discussed above are exact. They are not exact and do not adequately model random sequences which are constructed by shuffling residues (ibid). However, defining random sequences by sampling with replacement is not just convenient, it is more correct.

The way one constructs random sequences should represent the way sequences are constructed in nature. However, it is easier to describe how sequences diverge in nature. Ignoring the issue of

insertions and deletions, sequences diverge by point mutation. Of the two methods for constructing random sequences, this process of mutation best corresponds with sampling with replacement. One has little or no knowledge of which residues will mutate or what the resulting mutants will be. Yet, overall, one expects the set of mutant residues to reflect the average molar composition of proteins. On the other hand, constructing random sequences by sampling without replacement corresponds to a very different and unlikely view of how proteins diverge in nature. In this case, the DNA sequence fragments into pieces whose sizes are, on average, multiples of 3. These pieces are then randomized and ligated in such a way as to reassemble all the constituents and exclude any foreign pieces of DNA. Since the fragment sizes were multiples of 3, the reading frame and exact molar composition of the sequence is maintained. Clearly, a null hypothesis based on sampling with replacement is preferable to one based on the shuffling of sequences.

It is expected that the current null hypothesis will need to be modified by the results obtained with it. This hypothesis assumes the statistical independence of residues, meaning that knowledge of one portion of a sequence contains no information about the rest of the sequence. However, this is not always true. For example, endoduplications have been observed in many sequences and these patterns of repeated sequences can be very elaborate as in the case of fibronectin (Petersen et al., 1983; Tamkun et al., 1984). Sensitive tests of evolutionary relationships between proteins composed of arrays of repeated

patterns will require modification of the null hypothesis. Endoduplication can be thought of as the transposition of a part of a sequence into itself. These transpositions can also occur between different sequences and provide a further complication. Other instances of the non-independence of residues could be cited or imagined. However, such instances must first be characterized with the results obtained with the assumption of independence.

The second assumption I have made about sequences in this analysis is that they all share a common molar composition. However, this assumption is only convenient and not central to the method. One could use the actual molar composition of proteins in this analysis, for example.

1.5 Analysis of observational studies: Regardless of which program is used to compare sequences, a value for P(sim) can be calculated for each alignment. This value relates the quality of matching observed in the alignment, expressed in units of standard deviation, with the frequency of alignments expected from a population of unrelated sequences. However, this value is a measure of similarity and not kinship. The probability of kinship is dependent on both the value of P(sim) and the size of the database searched. This probability value, to be called P(kin), can be approximated as a one-tailed binomial probability:

$$P(kin) = \sum_{i=m}^{n} (p)^{i}(q)^{(n-i)} \binom{n}{m} \qquad \text{Eq. 1.1}$$

where n is the number of sequences in the database relevant to the comparison, m is the number of alignments whose P(sim) values are less than or equal to p, and q is (1-p). This approximation should be valid for the tail of any distribution of values for P(sim).

In Section 1.3, the population considered was the entire database of 2200 different sequences. This population is the appropriate object of comparison if one knows nothing more about the query sequence other than the sequence itself. However, more is generally known about the query sequence and this information can increase the sensitivity of a homology search by reducing the size of the database to include only those sequences that are relevant to the comparison. As an example, it might be known from a cloning strategy that a sequence is a eukaryotic extracellular protein. While this sequence ought to be compared with the entire database for the sake of completeness, it also ought to be compared with just the set of eukaryotic extracellular proteins. Additional knowledge about the sequence serves to reduce the value for n in Eq. 1. In general, the sensitivity of a comparison increases proportionally with the decrease in the number of relevant sequences searched. This increase in sensitivity can be sufficient to allow useful analogies to be drawn between the query sequence and a better characterized sequence.

A method for assessing similarity in which P(sim) is normally distributed, as the method in Section 1.4 appears to be, provides a more sensitive measure of homology. First, the signal to noise ratio of this method should be less than other methods

when the tails of their respective distributions are analyzed, as illustrated in Fig. 1.2. More importantly, the definition of P(kin) in Eq. 1 is strictly true throughout the distribution and this fact provides a powerful test of kinship. Consider a sequence which is not significantly homologous with any other sequence. Compare that sequence with a defined population such as all sequences of known ATP-binding proteins. If the query sequence is unrelated to ATP-binding proteins, then the resulting values of P(kin) should be normally distributed with a mean of 0.5. However, if the query sequence is related, then the values of P(kin) will deviate from this expected distribution. As an example, suppose there are 30 non-homologous ATP-binding proteins in the database and 15 of these are found to have values of P(sim) less than or equal to 0.1 when compared with the query sequence. The value for P(kin) by Eq. 1 for this observation is 4×10^{-8}, indicating kinship between the query sequence and a subset of ATP-binding proteins. Therefore, while no single alignment in this comparison reflects significant homology, kinship can be established.

Kinship will be distinguished from close homology in this discussion by the values for P(sim) describing the alignments. Define closely homologous sequences as those for which P(sim) is less than 10^{-5}. In this case, the relatedness of any two sequences could be demonstrated in a search of the entire database, given its current size. In contrast, kinship can be identified only within a defined population; no pair of sequences need be closely homologous. Kinship, then, is a quantifiable

attribute of a group of sequences, which may prove to be predictive of function. However, the frequency with which kinship correlates with function remains to be determined.

The analysis of kinship is not limited to comparing one sequence with a defined population; two populations can be compared as well. For example, the set of ATP-binding proteins can be compared with itself. If this set were composed of just five unrelated sequences (A, B, C, D, E), the resulting values of P(sim) might be:

A	-				
B	.98	-			
C	.61	.43	-		
D	.71	.81	.38	-	
E	.79	.53	.03	.18	-
	A	B	C	D	E

This set of scores is adequately described by a normal distribution and no kinship group can be identified within this set. In contrast, if the set contained some related sequences, the result might be:

A	-				
B	.01	-			
C	.04	.05	-		
D	.61	.31	.41	-	
E	.12	.30	.76	10^{-7}	-
	A	B	C	D	E

Here, sequences E and D are clearly homologous. In addition, sequences A, B, and C, with P(kin)= 0.008, would appear to constitute a separate kinship group.

Two different defined populations can be compared in the same way. For example, the set of ATP-binding proteins could be

compared with the set of proteases, these sets are presumably unrelated. The same expectation for unrelated sequences pertains as before: the distribution of P(sim) values should be normal with a mean of 0.5. Significant deviation from this distribution then implies kinship between members of the two groups. Kinship might be found, for example, between sequences of ATP-binding proteins (A - E) and NAD-binding proteins (V - Z):

A	.44	.25	.76	.27	.21
B	.79	.51	.36	.13	.77
C	.36	.43	.57	.33	.78
D	.01	.82	.02	.01	.06
E	.02	.53	.03	.01	.04
	V	W	X	Y	Z

For unrelated sequences, one expects that not only the overall distribution of scores be normal but that each row and column of scores is also normally distributed about a mean of 0.5. Given the data's deviation from those expectations, the conclusion from this comparison would be that the ATP-binding proteins D and E are related to the NAD-binding proteins V, X, Y, and Z.

There are two qualities which describe the alignments in a kinship group which will be called continuity and homogeneity. Continuity refers to the tendency of the regions of alignment to overlap with one another. Homogeneity is a property of overlapping regions of alignment such as that observed in the family of protein kinases and refers to the frequency with which similar amino acids occur at analogous positions. For example, when all tyrosine kinases are aligned, each sequence contains the invariant subsequence Gly-X-Gly at the same position in the alignment; this kinship group is homogeneous at this position.

One should be able to quantify each of these properties and calculate the probability of their observed values. When the probability values associated with either continuity or homogeneity are significantly low, then it is reasonable to postulate an ancient progenitor sequence. Techniques can then be developed for constructing a facsimile of this progenitor from existing sequences.

This strategy provides a potentially powerful method for determining the taxonomy of protein sequences. By analyzing only distribution tails, pairs and some families of homologous sequences can be identified. However, most sequences are not recognizably homologous with any other known sequence by this method. Yet by comparing defined populations of sequences, a greater number of taxonomic relationships should be apparent. Unless diverging sequences reach some cataclysmic point at which values for P(sim) go rapidly from approximately 10^{-4}, reflecting obvious homology, towards 0.5, some new relationships must be apparent.

It should be noted that while the above discussion has assumed a measure of similarity that is normally distributed, normality is not required for this analysis; it is just preferable. If P(sim) were not normally distributed, as in the case of the FASTP program, Eq. 1 would become progressively invalid the farther one moved from the tail of the distribution towards the mean. However, a valid cumulative distribution function could be found to describe this variable (McLachlan &

Boswell, 1985). Therefore, the same analysis could be performed although it would be more difficult.

A comprehensive program of study would compare all sequences in the database with one another, recording values of P(sim) for each pair of sequences. All obviously homologous pairs are identified at this step as those for which P(sim) is less than approximately 10^{-4}, given the current size of the database. Next all known functions of each sequence, or some manageable subset of these functions, are recorded in the sequence database. Here function is defined as ligand-binding activity, such as the binding of: ATP, NAD, actin, tubulin, etc. The sequences sharing a function are then compared as a group and kinship is assessed This is done for each defined function to determine the relatedness of sequences sharing a common function. The sequences sharing each function are then compared as a group with the sequences sharing every other function. In this way, the relatedness between groups of sequences mediating different functions can be determined.

The sensitivity of this approach can be tested with randomized sequences as discussed before (pg. 5). In this case, a mock phylogenetic tree of related sequences is constructed by progressive randomization. The structure of the tree is known beforehand and the program's ability to reconstruct it is tested. In addition, there are functions which are clearly similar, such as ATP- and NAD-binding. If this strategy is successful one expects to see corresponding similarities between the sequences of functionally similar proteins. Comparison of these sequences

would be of special interest because of the observed homology between the 3-dimensional structures of the two sets despite the absence of recognizable sequence homology. Other evidence of the evolutionary relatedness of nucleotide-binding proteins is available as discussed in Chapters 3 and 4. Therefore, the power of this approach can also be tested by the program's ability to duplicate these results.

Chapter II

Potential Application of Comparative Methods to Secondary Structure Predictions

The ability of proteins to renature to fully active conformations demonstrates that all the information required for determining correct three-dimensional structure is contained within the amino acid sequence (Anfinsen, 1967; Tanford, 1968). Much of this information acts in a local context by determining patterns of secondary structure in ways which are largely predictable, as discussed in Section 2.2. However, the consensus is that local sequence is not the sole determinant of secondary structure (Schulz et al., 1974; Matthews, 1975; Palau et al., 1982). Therefore, the prediction of secondary structure is presently an error-prone method. Because these errors tend to compound, predictions of overall tertiary structure are even more unreliable.

Given any error-prone method, measures of reliability are especially important. Since no method predicts each element of secondary structure with equal confidence, it is useful to quantitate levels of confidence for each. The methods in Chapter 1 can provide this desired measure of reliability as discussed in Section 2.3. It is expected that a substantial portion of the overall secondary structure predicted for any sequence will be found to be unreliable. However, those portions of structure

which are reliably predicted can serve as an additional basis of comparison between sequences. Moreover, when the structures of sets of evolutionarily related sequences are predicted, measures of reliability can be calculated for the entire set. If these sequences share a common structure, then this analysis should produce a consensus prediction the majority of which is reliable. Evolutionary relationships can be established by various criteria such as sequence homology or the pattern analyses discussed in Chapters 3 and 4.

2.1 The effects of protein sequence divergence on structure:

A number of studies have been done to determine how functionally homologous proteins retain those three-dimensional structures required for proper function despite divergence of their sequences. The observation common to each of these studies is that a majority of residues are found at the surfaces of proteins and are nonessential in the sense that they appear to mutate freely without destroying function. Interior residues, on the other hand, do appear to be constrained in a limited sense as discussed below. In contrast, residues directly involved in ligand binding tend either to remain unchanged or to have undergone conservative substitutions.

The earliest of these studies, by Lesk and Chothia (1980), was a comparison of the 3-dimensional structures of nine globins, including such divergent proteins as human alpha and beta hemoglobins, monomeric hemoglobin of the lamprey, myoglobin, insect erythrocruorin, and leghemoglobin (20). These globins are

heme-binding proteins with homologous tertiary structures. The elements of secondary structure are eight alpha helices, arranged in topographically identical patterns. The authors reported a number of important results:

1) The three-dimensional structures of the globins are mainly determined by approximately 59 residues which are involved either in the packing of helices or in the interactions between the helices and the heme.
2) The remaining 60 or so residues reside on the surfaces of the proteins and show very little sequence homology.
3) Thirty three of the 59 residues determining the structure of the proteins are buried. Mutations of these buried residues are conservative only in the sense that hydrophobicity is maintained. Within this constraint amino acid identities can vary. Volumes of side chains are not a constraint, as the mean change in volume is 56 $Å^3$, or roughly the size of a molecule of acetate. This represents variations in side-chain volumes of up to 57%.
4) The remaining 26 of these 59 residues are at the surface. They are apparently unconstrained even with respect to hydrophobicity.

The result of this variability in amino acid sequences is to shift the orientations of helices with respect to each other However, the number and topology of the helices remain constant. Despite the shifting of the helices, the geometries of the heme-binding sites of all the globins remain very similar. Moreover,

amino acids contacting the heme, particularly the heme iron, tend to be invariant.

Analyses similar to that done with globins were done with other protein families with similar results. In one study, 11 immunoglobulin domains were compared including both variable and constant domains of both heavy and light chains (Lesk & Chothia, 1982). The core structure of each of these domains is formed by 36 structurally homologous residues forming two beta-sheets which are packed face-to-face. In each of the domains, there is only one element of structural homology invariant in sequence: a disulfide bond bridging the two beta-sheets which has a tryptophan residue packed against it. The remaining residues lying between the beta-sheets were found to be only partially constrained with respect to acceptable mutations. As with the globins, hydrophobicity was the only identifiable constraining characteristic. The side-chain volumes of corresponding residues varied greatly with a mean variation of 60 $Å^3$. Maximum packing density of the crystalline-like hydrophobic interior was maintained by a number of compensatory mechanisms:

1) the sheets could be displaced relative to one other by rotating or tilting around the axis of the disulfide bridge.
2) hydrophobic side-chains from non-sheet residues could be packed between the β-pleated sheets.
3) hydrophilic β-sheet residues could be excluded from the hydrophobic interior by forming a beta-bulge.

These studies demonstrate that two aspects of protein structure remain conserved in the course of amino acid sequence divergence. First, the best conserved residues are those constituting a ligand-binding site. Methods for comparing and predicting putative ligand-binding sites are discussed in Chapters 3 and 4. The second point is that elements of secondary structure are also conserved, despite the divergence of the sequences constituting those elements. This conservation suggests the importance of secondary structure to the overall process of protein folding.

2.2 Predictors of secondary structure: Numerous predictors have been proposed to correlate amino acid sequence with elements of secondary structure. These predictors are of two types, physicochemical and empirical. The basic observation underlying physicochemical approaches is that residues with hydrophobic side chains tend to be situated within the interior of proteins. The attempt is then made to correlate patterns of hydrophobic residues within the primary structure with elements of secondary structure. Empirical methods are more statistical in character. Working from a database of known structures, one calculates the likelihood that a particular amino acid occurs within some element of secondary structure.

While physicochemical methods generally focus on the hydrophobicity of sequences, a variety of partially conflicting scales of hydrophobicity have been proposed (Hildebrand, 1979; Hansch & Leo, 1979; Tanford, 1980; Nemethey et al., 1981;

Wolfenden et al.,1981; Hvidt, 1983). There is as yet no consensus on which hydrophobicity scale is most appropriate due mainly to controversy over how hydrophobicity should be measured. Nonetheless, amino acid sequences can be translated with one of these scales into linear plots of hydrophobicity (reviewed in Rose et al., 1985). Each residue has associated with it a hydrophobicity value, called <H>, which is the average of both itself and a few neighboring residues. The number of neighboring residues which are included in the average varies between methods, but is typically between 5 and 9. By averaging over 5 or more residues, random fluctuations in the hydrophobicity profile are suppressed and regions of the sequence which are either especially hydrophilic or hydrophobic can be identified. Hydrophobicity profiles can then be interpreted as a physicochemical representation of the protein sequence.

Hydrophobicity profiles are useful in predicting the location of turns in tertiary structures (Rose et al., 1985). Since the majority of turns are situated at the surface of proteins and exposed to solvent, stretches of 4 or more hydrophilic residues generally correspond to turns. These hydrophilic regions are easily recognized as local minima in the hydrophobicity profile. The prediction of turns is of interest because turns are frequently the site of important interactions between a protein and its external environment. Some examples of these interactions are: 1) phosphorylation of proteins (Smith & Griffin, 1978); 2) N- and O-glycosylation of proteins (Aubert et al., 1976); and 3) immunological recognition (reviewed in Rose et al., 1985).

While local minima of hydrophobicity profiles correspond with residues localized at the surface of proteins, local maxima correspond with residues buried within the interior of the structure. These maxima also tend to correspond with periodic elements of secondary structure such as α-helices and β-sheets. Although hydrophobicity profiles do not contain enough information to distinguish these different elements of structure, the known periodicities of these elements can be used to help identify them. Eisenberg et al. (1982) have developed a measure of the amphiphilicity of helices; ie. a value describing the segregation of hydrophobic residues to one surface of a helix. For example, helices in globular proteins tend to be amphiphilic, the surface of the helix which is interior is hydrophobic while the remaining surfaces tend to be hydrophilic. In contrast, helices spanning the plasma membrane tend to be uniformly hydrophobic. This measure of amphiphilicity is called the hydrophobic moment, $<\mu H>$, and is the magnitude of the vector sum of hydrophobicity values for each residue in a helix. By comparing values for both hydrophobicity and hydrophobic moment, ($<H>$, $<\mu H>$), the authors were able to distinguish between helices of globular proteins, helices spanning membranes, and "surface-active" helices such as those which promote cell lysis. This method is only useful for relatively long helices of 12 or more residues. With shorter helices, simple statistical fluctuations in the calculated parameters tend to obscure the signatures of real helices. Because statistical fluctuations are inherent to

the method, this approach is not likely to be useful for overall predictions of secondary structure.

Empirical approaches provide the most useful predictors of overall secondary structure and many have been proposed. The method proposed by Chou and Fasman (1978) is probably the most commonly used. However, a similar approach was adopted by Robson & Suzuki (1976). Their analysis of the problem is the most complete and will be the focus of this discussion. They analyzed the relationship between sequence and conformation in a group of 25 proteins of known structure. The basic statistics they gathered were likelihoods that relate every amino acid to each secondary structure. For example, the likelihood that a glycine lies within a reverse turn is the ratio of the frequency of glycines within turns over the frequency of total residues within turns. The secondary structures considered were helices, reverse turns, extended conformations which include β-sheets, and random coils. Each residue of the 25 proteins was assigned to one of these classes on the basis of the structural data. Likelihood values were tested for their dependence on sample size and the majority were found to be convergent, meaning that these values should reflect general properties of amino acids. The reported likelihood values for the six rarest amino acids were only partially convergent and should be treated with caution.

Their method of structure prediction is presented in terms of information theory. In practice this simply means that they deal with the natural logarithm of likelihood values, called the information measure, rather than with the values themselves. When

predicting a type of structure, such as helices, a sequence is considered in segments of four residues and the information measures of those residues are summed. Progressing from one end of the sequence, these sums are calculated at each position of the sequence. When a sum is above some critical value, the segment is predicted to adopt the particular conformation in question. This process is repeated for each of the four types of structure, to yield an overall prediction of structure. This is the simpler version of the Robson method and is very similar to the Chou and Fasman method.

The extended Robson method involves a concept of "directional information". The authors found that the identity of a residue is predictive not only of its own conformation but of the conformations of neighboring residues. For example, the immediate neighbors of glycine residues are most likely to adopt the reverse turn conformation, which is not surprising. However, more distant neighboring residues are more likely to adopt a β-sheet conformation than a helical conformation. Directional effects can be classified into four main categories:

1) residues tending to form a particular conformation
2) residues tending to disrupt a conformation
3) residues tending to initiate a conformation, ie. occurring at the N-terminus of the structure
4) residues tending to terminate a conformation, ie. occurring at the C-terminus of the structure

This directional information is expressed, as before, as logarithms of likelihood ratios. Briefly, the extended method

augments the prediction produced by the simpler method by the addition of these new values.

As stated before, structure prediction is an error-prone method. The most accurately predicted element of secondary structure is the helix. With the Robson method, accuracies of about 75% of residues correctly assigned as either helical or non-helical are obtained. Similar accuracies are observed with the Chou & Fasman method (1978). Other elements of structure are predicted less accurately. There are two types of sources for this error. The first is that structure predictors are a statistical probe of protein structure and are necessarily imprecise. This fact may be sufficient in itself to account for most of the errors made by these methods. The second potential source of error is that interactions between residues distant in the sequence but juxtaposed in the tertiary structure may override the local effects that determine structure, the effects that these methods attempt to predict.

2.3 Measures of reliability: The simpler version of the Robson method is virtually identical to the FASTP method for scoring homology described in Section 1.2. In each case, the scores are the sums of log(likelihood) values. In homology searches, the score reflects the likelihood that one sequence is evolutionarily related to a second sequence. However, it was pointed out in Sections 1.3 and 1.4 that the probability of obtaining some score, rather than the score itself, was the value of interest in assessing similarity. This value, P(score), can be calculated for

any scoring system as discussed in Section 1.4. This treatment of homology eventually led to the variable P(kin) which describes the relatedness between sequences.

A similar approach can be applied to the prediction of secondary structure. In this case, the score for some segment of a sequence reflects the likelihood that that segment adopts a particular conformation. However, these scores are subject to statistical fluctuations and the score does not in itself describe the probability that the segment adopts the specified conformation. Values of P(score) can be calculated for the assignment of any segment to any conformational state using the methods of McLachlan & Boswell (1985) as discussed in Section 1.4. Therefore, P(score) provides a measure of the probability that a segment adopts the predicted structure. In the context of structure prediction, P(score) can be interpreted as a measure of the reliability of the prediction. For sequence segments which do not adopt the conformation in question the expected values for P(score) should be 0.5, on average. The smaller the values of P(score), the more probable it is that a segment adopts the conformation: ie. the prediction is more reliable.

Values of P(score) are a conservative measure of reliability in that predictions might be more reliable than these values indicate. P(score) is the probability that a random sequence, which is unrelated to the actual test sequence, will have a score as good or better than the test sequence. When searching for homology, the relationship being tested is an evolutionary one and a random sequence is clearly unrelated to any actual

sequence. With structure predictions, however, the relationship being tested is between sequence and conformation. If random sequences are actually constructed, they will fold too, and will contain the various elements of secondary structure. Therefore, many of the random permutations of sequence contributing to a value of P(score) will themselves adopt the conformation in question. Therefore, the actual reliability of a prediction should be better than that indicated by P(score).

When predicting the structure of a single sequence, this refinement in methodology is not expected to significantly improve the performance of the predictor. However, when these methods are applied to groups of related sequences, one should be able to predict consensus structures for the group which are highly reliable. The approach is similar to that in Section 1.4 for determining kinship between groups of sequences. As an example, consider the prediction of an α-helix in a set of sequences which are related and aligned by some independent criterion. Suppose each sequence is predicted to have a helix at analogous positions of the alignment, but that the values of P(score) for each prediction are only partially reliable with values between 0.2 and 0.1. If the set contained 5 sequences and all five were predicted to have a helix, ie. P(score) less than 0.2, the predicted consensus structure would be quite reliable. By the binomial formula, Eq. 1.1, the reliability is less than or equal to 3.2×10^{-4}.

There will be some complications to the prediction of overall consensus structures which will have to be dealt with for

this approach to be useful. The first is the effect of the independent criterion of relatedness on the prediction of structure. For example, homology is a natural criterion to use in establishing a relationship and an alignment between sequences. Here the effect of the homologies themselves on the prediction of structure must be accounted for. Second, methods must be developed for distinguishing between the most probable conformations that a segment might adopt: the same segment may appear to be either one structure or another of the four conformations studied. Third, elements of structure may not be perfectly aligned in the alignment of the sequences and methods are needed to determine whether these displaced structures are related.

This approach offers the possibility of predicting consensus structures for families of related sequences such as the family of tyrosine kinases. There are other families which will provide an immediate test of this approach. These are families for which crystallographic data is available such as serine proteases and globins. If the complications discussed above can be dealt with, the expectation is that the consensus structures derived from these families will be an accurate reflection of the known crystal structures. If they are not, the implication is that interactions between residues distant in the sequence but juxtaposed in the tertiary structure are a major determinant of protein folding.

Chapter III

Analysis of the Sequences of ATP-binding Proteins

This chapter describes two consensus patterns which are predictive of ATP binding and were derived by comparing ATP-binding sequences with structural data relevant to ligand binding. This analysis differs from the methods in Chapter 1 by its reliance on structural data which allows the identification of distant evolutionary relationships. However, the two approaches complement each other.

Section 1.5 introduced methods of kinship analysis as a means of increasing the sensitivity of sequence comparisons. Kinship analysis is performed on groups of sequences sharing a common function and tests whether those sequences share enough similarity to support a hypothesis of common ancestry. Ideally, one would like a physical interpretation of these results. When the alignments of these sequences are examined they should share a common region of homology which mediates the function they share. Moreover, certain residues might be particularly conserved among the sequences allowing the definition of a consensus homology. The discussion of protein divergence in Section 2 1 made the point that residues contacting a ligand tend to be the best conserved residues in a protein. This conservation was observed in the ligand-binding sites of globins (Chothia and Lesk, 1976), cytochromes c (Chothia and Lesk, 1982), and azurins

and plastocyanins (Lesk and Chothia, 1984). The expectation, then, is that the residues of a consensus homology should include those which contact the ligand defining the functional group. As a result, the consensus homology should be a predictor of the ligand-binding activity shared by the functional group. Moreover, one should find that structural data describing a ligand-binding site for one protein is generally relevant to the other members of the kinship group.

My work with sequence comparison began with the derivation of consensus patterns predictive of specific functions, focussing on nucleotide-binding activity. One result of this work is a method for the statistical analysis of matches to a consensus pattern which can be applied to kinship groups. However, the emphasis of this study was opposite to that of kinship analysis. The working hypothesis was, from the start, that structural data for one protein were relevant to proteins sharing the same function and groups of sequences were compared with that data. The aim in these studies was to identify sequence elements involved in ligand binding which are common to large classes of proteins. Consequently, comparisons were done with sequences which shared no homology detectable by pairwise comparisons since such homologies would tend to obscure the identification of these essential elements.

The comparisons described in this chapter are of ATP-binding sequences and result in the definition of two consensus patterns, ATP-1 and ATP-2, which are demonstrated to be predictive of that function. The patterns consist of elements whose positions are

fixed relative to one another and could be thought of as residues in consensus homologies. However, the pattern elements are small and can be scattered through the protein sequences. In the case of ATP-1, the pattern spans 153 residues but consists of just 16 residues arranged in 11 elements; no element is adjacent to any other element. The complexity of the ATP-1 pattern strongly suggests evolutionary relationships between the sequences containing it. However, it is unclear whether any method of sequence comparison which does not also rely on structural data could identify this relationship. In contrast, the elements of the ATP-2 pattern are much more clustered and it is likely that this pattern could be identified by an analysis of kinship.

The utility of any method of sequence comparison is dependent on the extent to which proteins have tended to evolve from pre-existing forms as opposed to having arisen independently. Among ATP-binding sequences, the evolutionary origin of protein diversity is clear. Of all sequenced ATP-binding proteins, 41% contain significant matches to the ATP-1 pattern and 33% contain matches to the ATP-2 pattern. Similar evolutionary relationships can also be demonstrated between other nucleotide-binding proteins as discussed in Chapter 4. Unless nucleotide-binding proteins are atypical, these results suggest that much of protein phylogeny can be traced by comparing groups of sequences sharing a function.

3.1 The ATP-1 pattern: The analysis proceeds in a series of steps. The first is a comparison of sequences of ATP-binding

proteins with structural data on the binding of the related nucleotide NAD. This comparison yields an initial pattern. ATP-binding proteins containing this pattern are aligned by it and positions in the alignment are searched for similar amino acids occurring in a significant fraction of the sequences. This second comparison produces a list of similarities which is compared with additional structural data to identify other sequence elements characteristic of ATP binding. These elements together with the elements of the initial pattern become the consensus pattern. A numerical measure of matching is established which allows straightforward tests of the correlation between pattern and function. The results of these tests justify the use of the consensus pattern as a predictor of ATP-binding activity and a simple method of making such predictions is described. The approach should be applicable to other ligand-binding functions.

3.1.1 Derivation of the pattern: The analysis began with the comparison of the sequences of ATP-binding proteins with that of lactate dehydrogenase (LDH). LDH is an NAD-dependent enzyme whose three-dimensional structure, including details of the NAD-binding site, is known (Grau, Trommer, & Rossmann, 1981; Eventoff et al., 1977). The NAD-binding domains of LDH and several other dehydrogenases are very similar to one another in that their three-dimensional structures are largely superposable (Ohlsson et al., 1974). Moreover, those portions of the binding domains accommodating the ADP moiety of NAD are also very similar to the structures of ATP-binding domains by the same criteria of

structural homology (Rossmann et al., 1974; Schultz et al., 1974; Rossmann et al., 1977). While the sequences of these proteins were not obviously homologous by pairwise comparisons (Ohlsson et al., 1974; Rossmann et al., 1977), the observed structural homologies suggested that the sequences of these nucleotide-binding sites might be related nonetheless.

The sequences of four ATP-binding proteins (actin, cAMP-dependent protein kinase, v-mos, and asparagine synthetase) were searched visually for sequence elements similar in both identity and position to those elements in LDH which constitute the ADP portion of the NAD-binding site. While the identities and approximate positions of the five elements were determined by this comparison, the final positions within which the elements were sought were determined in another procedure. The set of five elements was used as a pattern in searches of the sequence database (Barker et al., 1983). The spacings of the elements were adjusted to increase the yield of ATP-binding proteins produced by the search. The final version of the pattern generated by this procedure is shown in Fig. 3.1 and differed from the initial pattern only in the addition of a few residue positions to the spacings of elements A3, A4, and A5. The chosen pattern of elements in the ATP-binding proteins differed from that in LDH by the insertion of approximately 15 residues between elements A3 and A4.

As described in Fig. 3.1, it was found that 14 of 30 sequences of ATP-binding proteins in the 1983 sequence database contained matches to the pattern of five elements. This initial

Fig. 3.1: Initial pattern of sequence elements common to many ATP-binding proteins.

a) the sets of functionally similar amino acids used in the analysis and the symbols designating them.

b) a pattern of sequence elements characteristic of many ATP-binding proteins which was derived by comparisons with structural data on the binding of the ADP moiety of NAD by lactate dehydrogenase. Residues analogous to those contacting the ADP moiety are circled. Fourteen of the 30 ATP-binding proteins in the 1983 database contain elements A1-A4 and 7 of those contain element A5 as well. These include proteins which bind ATP (or AMP) as an effector: amidophosphoribosyl transferase (Tso et al., 1982); glycogen phosphorylase (Stura et al., 1983); transferrin (Egyed et al., 1975); and proteins which are ATP-dependent: β subunit of RNA polymerase; β subunit of F1 ATPase (Khananshvili et al., 1985); SV40 large T antigen (Clertant et al., 1984); asparagine synthetase; alanyl-tRNA synthetase; actin; T7 DNA ligase (Kornberg, 1980); primase (Kornberg, 1980); cAMP-dependent protein kinase; v-mos (Kloetzer et al., 1983); v-fps. The last three proteins are protein kinases and are known to share sequence homology (Lipman & Pearson, 1985) but aside from elements A1 and A2, this homology does not correspond with pattern elements.

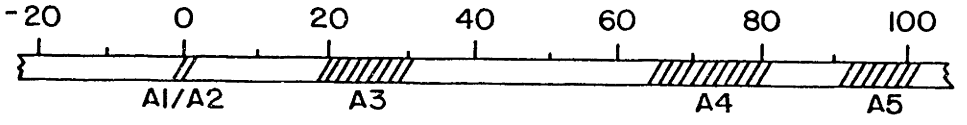
A.

+ = .cationic residues = {H,K,R}

− = anionic residues = {D,E}

@ = aromatic residues = {F,H,W,Y}

B.



A1 Ⓖ at 0

A2 Ⓖ at 2

A3 (−..+) or (− ..+), − is 20-31 residues from A1

A4 (−@) or (@−), @ is 65-80 residues from A1

A5 (⊕..−), + is 10-20 residues from A4

pattern was used to align the 14 matching sequences for the next step of the analysis. The subsequence Gly-x-Gly (elements A1 and A2) was chosen as the point of alignment for the sequences because it was the most conserved element, being found at structurally homologous positions in the nucleotide-binding domains of most dehydrogenases studied and also of many ATP-binding proteins (Ohlsson et al., 1974; Schultz et al., 1974).

A program was written to search systematically for similarities between the aligned sequences, ie. for ranges of positions within which similar amino acids are observed to occur among a significant fraction of the aligned sequences. Twenty-three equivalent classes of amino acids were considered including the identical matches corresponding to the 20 amino acids and the three sets of conservative substitutions shown in Fig. 3.1. While similarities were identified by low probability values, calculated as shown in the legend of Table 3.1, these values do not in themselves indicate a significant relationship between the sequences. This is because the similarity program makes thousands of individual comparisons and even a random collection of sequences produces a set of ostensibly significant similarities.

The issue was whether these similarities corresponded with residues known to be involved in the binding of ATP. This was accomplished by comparing the similarities with available structural data, as shown in Table 3.1. The structural data used for these comparisons included crystallographic studies detailing 1) the binding of ATP to adenylate kinase (Pai et al., 1977) and 2) the binding of AMP to the effector site of glycogen

Table 3.1: Comparison of structural data with similarities observed in the sequences of 14 ATP-binding proteins.

@ The comparison of the 14 sequences produced a list of 30 similarities with probability values less than 0.01. The table lists only those similarities analogous to the sequence elements in column 1.

¶ The expected frequencies for similarities were calculated with the binomial formula:

$$freq = 1 - (1 - mf)^x$$

where mf is the molar fraction of an amino acid observed in the sequences analyzed and x is the number of positions within which the amino acid may occur. Each residue was assumed to be statistically independent of other residues.

The fraction of sequences containing an analogous similarity excludes the sequence from which structural data was obtained.

† The P-value for the similarity is calculated with the binomial formula:

$$P = \sum_{i=m}^{n} (freq)^i (freq)^{(n-1)} \binom{n}{i}$$

where freq is the frequency for the similarity, m is the number of sequences containing the similarity, and n is the total number of sequences in the alignment.

Table 3.1: Comparison of structural data with similarities

* Residues involved in ligand binding	@ Similarities analogous with ligand-binding residues	¶ Frequency calculated for similarity	# Fraction of sequences containing similarity	† P-value
	- Adenylate kinase -			
H at 15	H at 9-26	0.33	11/14	6.5×10^{-4}
V at 47 V at 54	V at 47-54	0.44	12/14	1.8×10^{-3}
L at 53				
L at 56				
I at 72				
R at 77				
Q at 81	Q at 72-82	0.36	10/14	8.6×10^{-3}
	- Glycogen phosphorylase -			
Y at -1				
R at 37				
P at 38				
E at 39	- at 39	0.13	6/13	6.1×10^{-3}
K at 91	K at 91-94	0.18	7/13	6.9×10^{-3}
D at 150				
R at 152	+ at 152	0.13	6/13	6.8×10^{-3}
S at 156				
	- cAMP-dependent protein kinase -			
K at 22	K at 19-22	0.18	6/13	1.8×10^{-2}

phosphorylase (Stura et al., 1983), along with 3) results of affinity labeling of cAMP-dependent protein kinase (Zoller et al., 1981). Two of these proteins were among the 14 aligned sequences. Adenylate kinase, while not one of the 14 sequences due to its lack of element A4, contained elements A1-A3 and could be aligned with the other sequences. The residues of these enzymes cited by the respective authors as being involved in ligand binding are listed. Similarities which are analogous with these residues are listed along with the probability values associated with them. Seven new elements were selected by this comparison of similarities with structural data and were added to the original five to form the canonical pattern shown in Table 3.2.

3.1.2 Assessing the significance of matches to the consensus pattern: The pattern in Table 3.2 is a consensus which a significant fraction of ATP-binding proteins are expected to resemble, though no sequence is necessarily expected to be identical with it. Two techniques are required to make a pattern useful: 1) a way to measure similarity between a pattern and any match observed at a given location in a sequence, and 2) a way to estimate the significance of observing the match anywhere in the sequence. Both techniques are required because the relevance of an observed match to the binding of ATP is dependent not only on the quality of the match but on the number of residues searched. With these techniques, the correlation between the canonical pattern and ATP-binding can be proven. The same sort of

Table 3.2: The ATP-1 consensus pattern

Element No.	Description	p	q	
A1	G at 0	0.074	0.926	
A2	G at 2	0.074	0.926	
A3	(-..+) or (-...+) at 20 to 31	0.285	0.715	
A4	(-@) or (@-) at 65 to 80	0.308	0.692	
A5	(+..-) at 10 to 20 from A4	0.151	0.849	initial elements
A6	K at 19 to 22	0.232	0.768	new elements
A7	- at 39	0.110	0.890	
A8	V at 47 to 54	0.421	0.579	
A9	Q at 72 to 82	0.352	0.648	
A10	K at 91 to 94	0.232	0.768	
A11	+ at 152	0.135	0.865	
A12	H at 9 to 26	0.351	0.649	

The pattern is composed of the initial pattern elements in Fig. 3.1 together with the elements resulting from the analysis in Table 3.1. Except where noted in the description of the elements, positions refer to C-terminal displacement of an element from element A1. Relative frequencies of elements, p for presence and q for absence of an element, were calculated from the average molar composition of a large set of protein sequences (Dayhoff et al., 1978), in the same manner as in Table 3.1. As a class, ATP-binding proteins did not deviate significantly from this average composition.

calculation provides an a priori estimation of the probability that a given sequence represents an ATP-binding protein.

Considering the pattern as a set of elements, the question of similarity becomes a matter of determining which subsets of the elements represent significant matches. If the elements were equiprobable, significance could be assessed with the binomial formula. Because they are not, the expansion underlying the binomial formula is calculated explicitly and summed to form a cumulative distribution function. Figure 3.2 plots these cumulative probabilities (S_x) versus the probabilities of specific patterns (P_x) and can be used to calculate a measure of similarity manually as described in the legend.

The resulting measure of similarity gives the probability of a match occurring at only one particular location in a sequence. The probability of a match occurring anywhere in the sequence could be easily calculated if each position in that sequence could be treated as an independent trial for the presence of a match. Because the values of interest for S_x are small, such independence would imply that the occurrences of significant matches follow a Poisson distribution (Sokal & Rohlf, 1969). While sequence positions are not strictly independent trials, it was found in searches of a series of very long randomly generated sequences that the occurrences of matches are adequately modeled by the Poisson distribution. Thus the mean of the numbers of matches observed per sequence was equal to the product of S_x times the number of residues searched and also equal to the

Fig. 3.2: Cumulative distribution function of pattern matches.

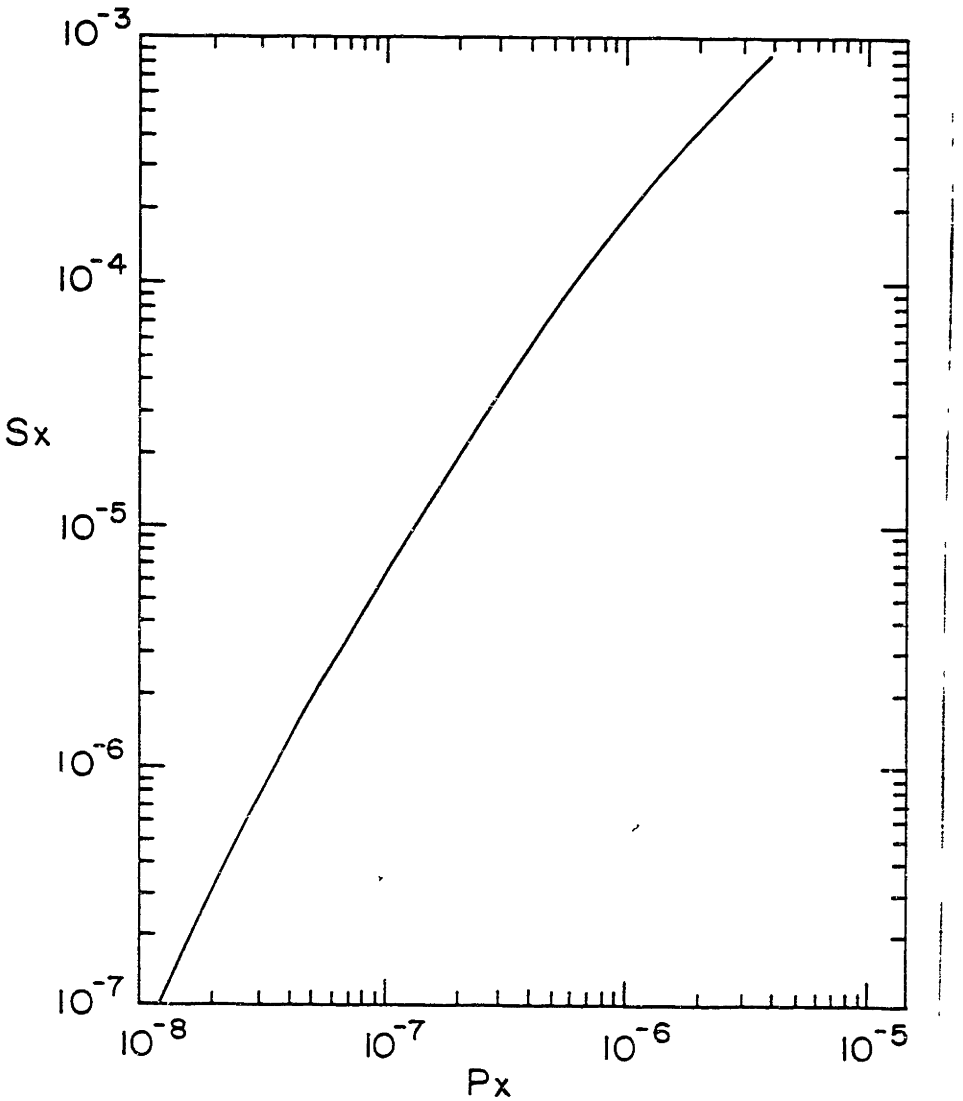
The probability of observing any particular pattern by chance is the product of the values for p of every element in Table 3.2 which is present times the product of the values for q of every element absent. Therefore the probability of observing a perfect match is:

$$p_1 \times p_2 \times p_3 \times p_4 \times p_5 \times p_6 \times p_7 \times p_8 \times p_9 \times p_{10} \times p_{11} \times p_{12}$$

and the probability of some less than perfect match would be, for example:

$$p_1 \times p_2 \times q_3 \times q_4 \times p_5 \times p_6 \times q_7 \times p_8 \times q_9 \times q_{10} \times p_{11} \times p_{12}$$

The relevant probability value for a match is given by the probability of observing precisely that match (P_x) plus the probabilities of every match as good or better. This sum of probability values (S_x) is a measure of the quality of the match. The plot relates values of P_x, which are easily calculated, to S_x.


10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
Sx
10^{-8}
10^{-7}
10^{-6}
10^{-5}
Px

variance of these observed matches. However with shorter sequences, the occurrences of matches deviated from the Poisson distribution. This is because the pattern spans 153 residues and the C-termini are necessarily lacking in pattern matches. It was found, again with randomized sequences, that the last 95 residues of C-termini are virtually devoid of significant matches and that subtracting 95 from the number of residues searched is an adequate correction for this end effect.

3.1.3 Correlation of the pattern with ATP binding: The correlation of the pattern with the binding of ATP was tested as described in Figure 3.3. Figures 3.3a and 3.3b show that both for randomly generated sequences and for the sequences of proteins not known to bind ATP, the number of pattern matches observed is that expected by chance. In contrast, two independent sets of sequences of ATP-binding proteins contain matches with a frequency higher than expected by chance, as shown in Fig. 3.3c and 3.3d. The P-values for the deviations from the expected frequency of matching when $S_x \leq 3 \times 10^{-4}$ were 0.035 for the sequences in Fig. 3.3c and 0.021 for those in Fig. 3.3d. Thus in two independent tests of the pattern performed with sequences not used in the derivation of the pattern, P-values less than 0.05 were obtained proving the correlation of the pattern with the binding of ATP. Including sequences used in the pattern's derivation, 41% (24 of 58) of the ATP-binding proteins sequenced to date contain a significant match to this pattern.

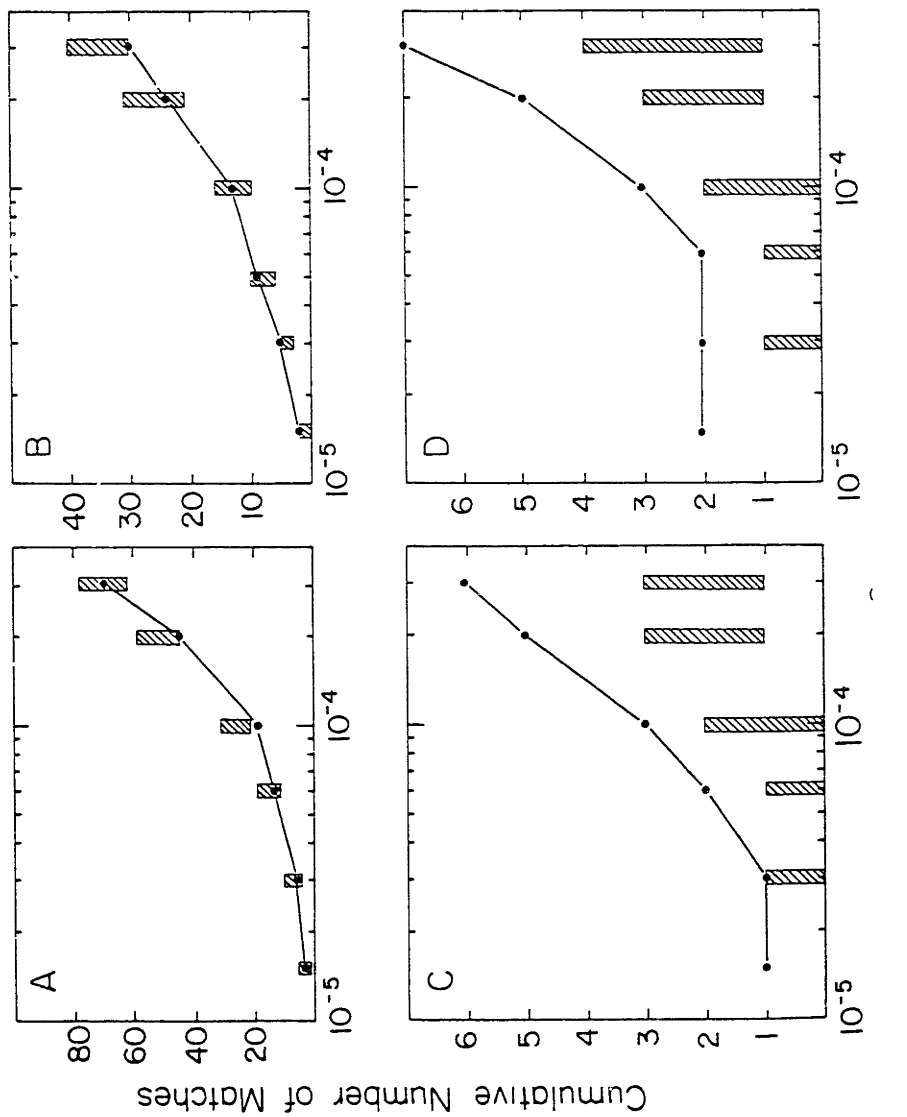
Figure 3.3: Comparison of cumulative numbers of observed pattern matches with numbers of matches expected by chance.

Each graph plots the cumulative number of matches observed at given values of Sx. Cross-hatchings represent the ranges of the cumulative numbers of matches expected by chance for a given value of Sx.

a. Matching observed in randomly generated sequences.

b. Matching observed in the sequences of proteins not known to bind ATP.

c. and d. Matching observed in two sets of sequences of known ATP-binding proteins. The P-value for the deviation from the expected frequency of matching calculated with the binomial formula when ($Sx \leq 0.0003$) is 0.035 in c and 0.021 d, demonstrating the correlation of pattern with function. The sequences of ATP-binding proteins with pattern matches were: in c) glutamate dehydrogenase (Pal et al., 1979), dnaA protein (Kornberg, 1980), phosphoglycerate kinase, and phage T7 DNA primase, T7 protein kinase and T7 DNA ligase (Kornberg, 1980); in d) arginino-succinate synthetase, fumarase (Miles & Guest, 1984), hygromycin B phosphotransferase (Gritz & Davies, 1983), carbamoyl-phosphate synthetase, galactose kinase, γ-glutamyl kinase, γ subunit of phosphorylase kinase, and nematode myosin. The match observed in T7 DNA ligase was at a different location in the sequence than its match to the initial pattern making this observation independent of the derivation of the pattern.
Methods: Random sequences were generated by sampling with replacement from a pool of residues reflecting the average molar composition of proteins (Dayhoff, et al., 1978). Sequences of proteins not known to bind ATP (b) and one set of known ATP-binding proteins (c) were from the 1985 protein sequence database. A second set of sequences of ATP-binding proteins (d) came from a search of recent literature. Both sets of sequences of ATP-binding proteins exclude those 14 sequences used to generate the pattern. Allowing for the end correction described in the text, the number of residues in each set which could be effectively searched for the pattern were: a) 228,582 residues in 459 sequences; b) 131,390 in 400 sequences; c) 8,831 residues in 32 sequences; and d) 10,057 residues in 10 sequences. Multiple members of homologous families of sequences, such as immunoglobulins, were not included in the count of residues. For each value of Sx, the number of matches expected by chance is $nS_x \pm \sqrt{nS_x}$; where n is the number of sequences searched.


A
B
C
D
Cumulative Number of Matches
10^{-5}
10^{-4}

3 1 4 Predicting ATP binding by inspection: The fact that proteins which do not bind ATP behave as a set of random sequences with respect to this pattern suggests that the a priori probability of a match to the pattern can be used as a direct predictor of ATP-binding. A program was written which searches sequences for matches to any specified pattern. The program generates a cumulative distribution function for the pattern and tests the significance of each pattern match. For the ATP-1 pattern, the information in Table 3 2 and Fig. 3 2 is sufficient for the prediction of ATP binding by inspection, as described below.

Because the great majority of pattern matches contain both elements A1 and A2, reasonably thorough searches can be made by looking first for the subsequence Gly-x-Gly and then looking for the other elements in Table 3.2. For any observed match, calculate P_x as the product of all values for p of elements present times all values for q of elements absent. Next, find the value for S_x corresponding to P_x from Fig. 3.2 The probability, C, that a match represents ATP-binding activity is a function not only of the quality of the match but also of the number of residues searched:

$$C = (1 - S_x)^{(n-95)}$$

where n is the number of residues in the sequence searched and is corrected for the end effect by subtracting 95. The term n refers to all of the sequences searched; if for example, the sequence of

a virus is searched and one of the coding sequences contains a pattern match, the sum of all coding regions is the relevant value for n, not the size of the one coding sequence containing the match.

The prediction that gene 46 of phage T4 binds ATP provides an example of the method The sequence of this gene (Gram & Rueger, 1985) contains a match consisting of the elements A1,2,3,4,7,9,10,11. Calculating from the values in Table 3.2, the value of P_x for precisely this match is 1.43×10^{-7}, and from Fig. 3.2 the value for S_x is 1.0×10^{-5}. Gene 46 is composed of 560 residues but was sequenced along with another 8 genes comprising an additional 1620 residues. The significance of the match is dependent on the total number of residues searched This number is 1325 when 95 is subtracted from the number of residues in each of the 9 coding sequences to correct for the end effect Therefore the probability value for this match is 0.01 and gene 46 is predicted to bind ATP.

3 2 The ATP-2 pattern: Another predictor of ATP-binding was proposed by Walker et al. (1982) and has been subsequently observed in other ATP-binding proteins (Finch and Emmerson, 1984). By comparing sequences with the DIAGON program (Staden, 1982), Walker et al. identified a small region of sequence homology held in common by a number of ATP-binding proteins and derived a consensus sequence which seemed diagnostic. Comparing

Table 3.3: The ATP-2 consensus pattern

Element No.	Description	p	q	
B1	G at 0	0.074	0.926	
B2	K at 1	0.064	0.936	
B3	T at 2	0.061	0.939	
B4	I or V at 9	0 116	0.884	
B5	G or A at -5	0.156	0.844	
B6	L, I, or V at -7	0.204	0.796	
B7	+ at -12 to -11	0.251	0 749	original elements
B8	G at -2	0.074	0.926	new elements
B9	E at 25 to 29	0.260	0.740	

A second consensus pattern correlating with the binding of ATP This is a refinement of a pattern proposed by Walker et al (1982). Elements with negative values for positions are N-terminal to element B1 Relative frequencies of elements, p for presence and q for absence of an element, were calculated from the average molar composition of a large set of protein sequences (Dayhoff et al., 1978), in the same manner as in Table 3.1.

just the sequences they used in their analysis with available structural data resulted in the definition of two additional elements giving the pattern shown in Table 3.3. The additional Gly at -2 in the pattern was chosen by analogy with the NAD-binding sites of the dehydrogenases and with elements A1 and A2 of the pattern in Table 3 2; with element B1, it forms the typical Gly-x-Gly subsequence Three of the four sequences originally compared by Walker et al contained this glycine, an outcome with a P-value of 1.6×10^{-3} The second new element was B9 in Table 3.3. It was chosen because it corresponds to a site in the β subunit of F1 ATPase modified by dicyclohexylcarbodiimide only in the absence of ATP (Yoshida et al , 1981). While this is only weak evidence for that residue residing within the ATP-binding site of F1 ATPase, all four of the other sequences contained glutamate residues at comparable positions ($P= 1.1\times10^{-3}$).

This pattern was analyzed in the same way as the ATP-1 pattern. Fig. 3 4 is the graph of the cumulative distribution function for this pattern relating values of P_x with those of S_x The analyses of the frequencies of matching are shown in Fig. 3 5. Note that, as with the previous pattern, the frequency of matching observed in randomly generated sequences is within the expected ranges (Fig 3 5a) while the frequency of matching among known ATP-binding proteins is significantly different (Fig. 3.5c). The P-value calculated for this deviation at $S_x \leq 3\times10^{-4}$ was >0.01, confirming the validity of this pattern.

Figure 3 4: The cumulative distribution function for matches to the ATP-2 pattern

This graph for the distribution function for the ATP-2 pattern is analogous to that for the ATP-1 pattern in in Fig. 3.2.

Method for predicting ATP-binding by inspection: The method is analogous to that described in Sec. 3.1.4. Inspection of sequences can be simplified by searching first for the subsequence G-K. Because the pattern is shorter than that in Fig. 2, the probability that a match represents the binding of ATP is given by the equation:

$$C = 1 - (1 - S_x)^{(n - 20)}$$

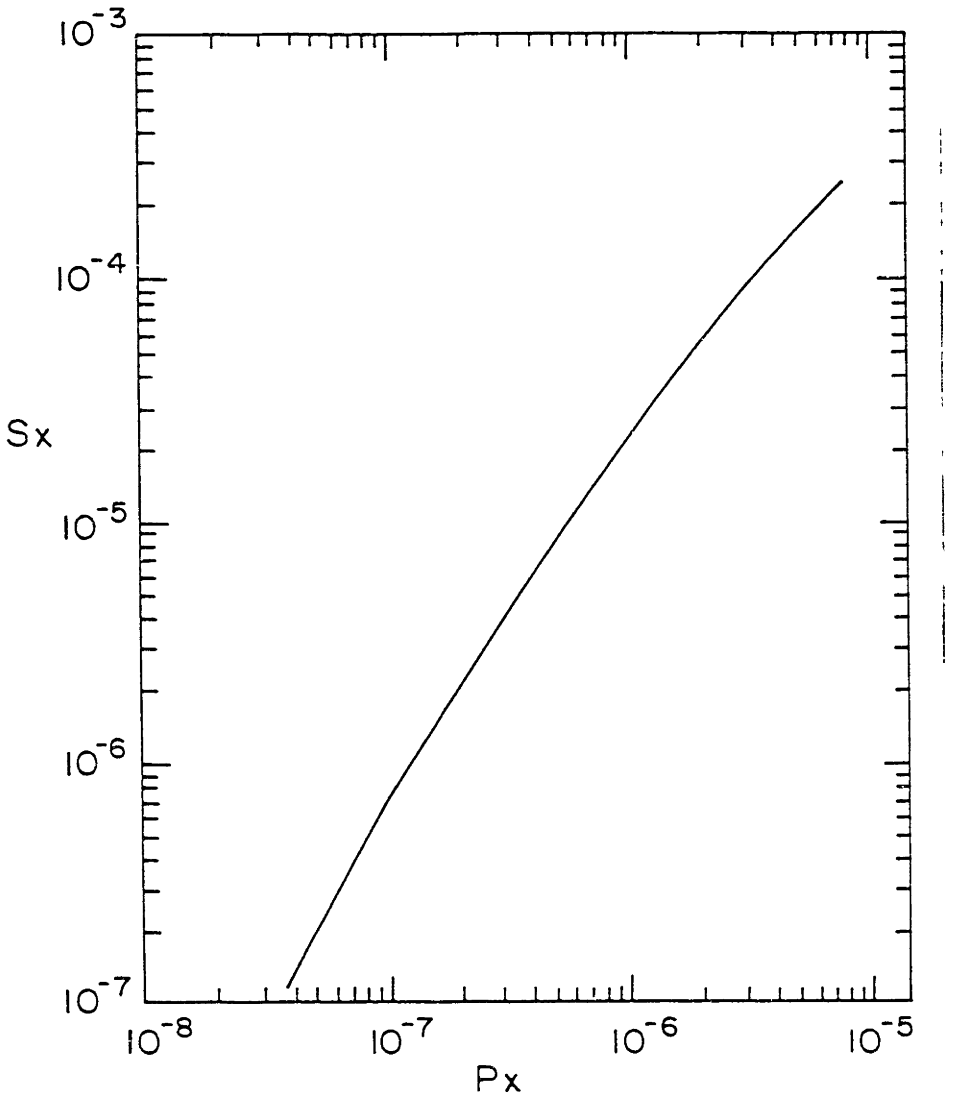

10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
Sx
10^{-8}
10^{-7}
10^{-6}
10^{-5}
Px

Figure 3.5 Test of the correlation of the ATP-2 pattern with the binding of ATP.

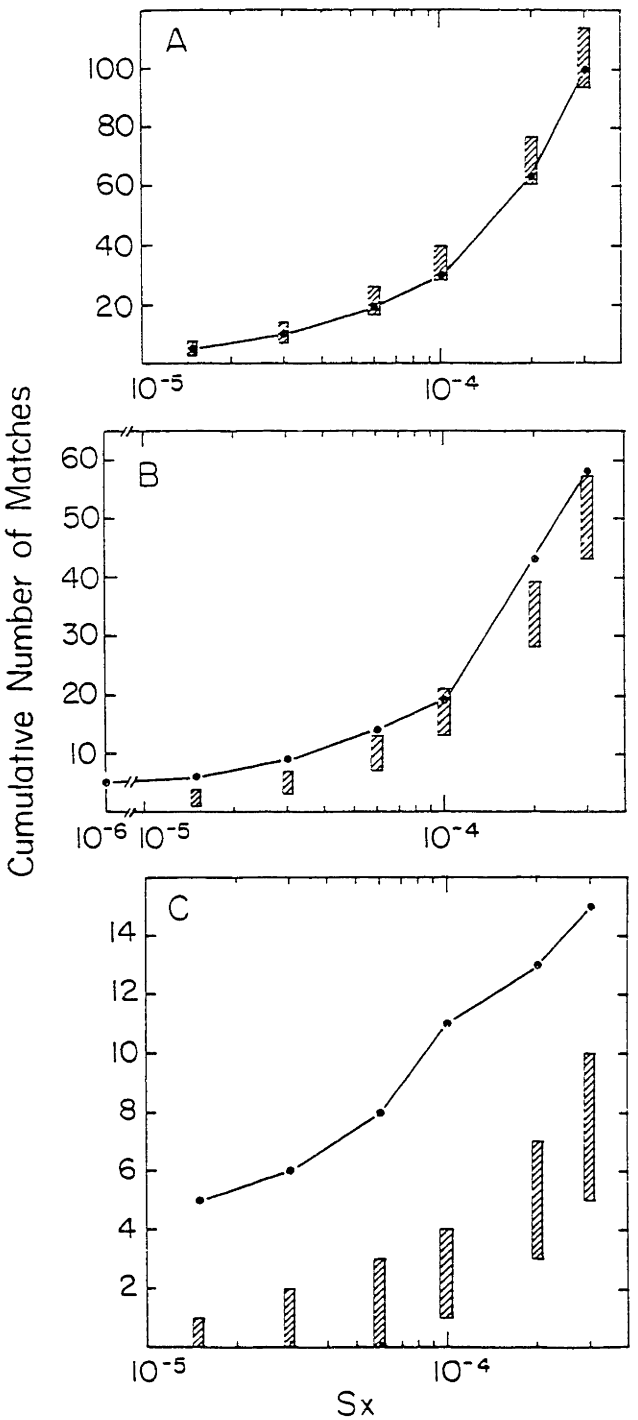
Each graph plots the cumulative number of matches observed in a set of sequences at given values of Sx. Cross-hatchings represent the ranges of numbers of matches expected by chance.

a. Matching observed in randomly generated sequences.

b. Matching observed in the sequences of proteins not known to bind ATP

c. Matching observed in sequences of known ATP-binding proteins; the P-value calculated for the deviation from the expected number of matches at ($Sx \leq 0\ 0003$) was >0 01 The sequences of ATP-binding proteins with matches to this pattern were· thymidine kinase, aspartokinase, the histidine permease inner membrane protein (Higgins et al., 1985), SV40 large T antigen, b subunit of RNA polymerase, ATP phosphoribosyl transferase, homoserine kinase, glutaminyl-tRNA synthetase, the oncogene src, actin, Fe protein of the nitrogenase complex (Mortenson & Thornley, 1979), and the gene products of uvrD, dnaA, dnaB, and rho (Kornberg, 1980)

Methods: The methods are the same as in Fig. 4.3. For each set of sequences, the number of residues which could be effectively searched were: a) 346,860, b) 165,815, and c) 24,738. The pool of sequences of ATP-binding proteins used in this test included those sequences matching the first pattern and excluded those 4 sequences used in generating the pattern in Table 4 3.


A
100
80
60
40
20
10^{-5}
10^{-4}
B
60
50
40
30
20
10
10^{-6}
10^{-5}
10^{-4}
C
14
12
10
8
6
4
2
10^{-5}
10^{-4}
Sx
Cumulative Number of Matches

However, in contrast with the results from the previous pattern, the set of proteins not known to bind ATP (Fig. 3.5b) deviated somewhat from the range of expected values This deviation can be ascribed to five sequences containing matches with conspicuously low values for S_x which range from 5 2×10^{-9} to 1.3×10^{-6}. These low values suggest that these five proteins bind ATP, they are 183 3K protein of tobacco mosaic virus and 125K protein of alfalfa mosaic virus (which share further homology in this region), phage fd gene I, phage T7 DNA maturase B, and lambda phage Ea59 gene.

Seven ATP-binding proteins contain matches to both patterns For actin and myosin, the two patterns predict the same location for the ATP-binding site However, the two patterns select different regions of the sequences of five of the seven proteins. These double matches cannot be reconciled with experimental data in the cases of the dnaA protein or the β subunit of E. coli RNA polymerase. Yet for three of these proteins, the separate matches reflect function since these proteins are known to have two distinct nucleotide-binding domains. These proteins are adenylate kinase (Pai et al., 1977), the β subunit of F1 ATPase (Khananshvili et al., 1985), and glutamate dehydrogenase (Pal and Colman, 1979).

3.3 The relationship of sequence pattern to protein structure:

Since the pattern elements were chosen by a direct comparison with residues known to constitute the ligand-binding sites of particular proteins, the majority of elements in a pattern match

ought to be constituents of the ATP-binding site The correlation of pattern elements with the structure of the ligand-binding site is further supported by the fact that the original elements A1, A2, and A3 are at the ATP-binding site of adenylate kinase and that elements A1, A2, and A4 are at the AMP-binding site of glycogen phosphorylase. Also, recent evidence suggests that element A11 of adenylate kinase is involved in the mechanism of transfer of the γ-phosphate of ATP to AMP (Fry et al., 1985). Since the element A11 originated from the comparison of sequences with structural data from phosphorylase, its relevance to the structure and function of adenylate kinase is remarkable.

If the patterns reflect the structure of ATP-binding sites, it follows that mutagenesis directed against pattern elements ought to alter the affinity of a protein for ATP which offers a way to test this interpretation of the patterns. Elements of special interest in this respect are those likely to be involved in the catalytic mechanism of ATP hydrolysis. Site-directed mutagenesis of these elements might have a very specific effect, abolishing the catalytic activity of a protein yet leaving it with a measurable affinity for ATP. In addition to element A11 discussed above, the histidine corresponding to element A12 has been implicated in the phosphate transfer mechanism of adenylate kinase (Pai et al., 1977; Cohn, et al., 1972). Another histidine at positions 110 to 120 of the pattern was found to occur in the sequences with a frequency comparable to element A12 (not shown) suggesting that histidines at this position might also be mechanistically important. The family of protein kinases,

including those oncogenes which phosphorylate tyrosines, contain matches to the ATP-1 pattern. Members of this family contain the putative catalytic residues and would be good candidates for site-directed mutagenesis.

The hypothesis that pattern elements contact ATP in the protein-ligand complex implies only that these residues must cluster around a region corresponding to the binding site. It would be interesting if more could be said of the structures of these proteins. The structural homologies observed between the few nucleotide-binding proteins that have been crystallized suggest that the majority of these proteins share a common form The methods of structure prediction discussed in Chapter 2 could be applied to the sequences sharing a pattern. By considering a multiple alignment of all proteins sharing a sequence pattern, it should be possible to generate a consensus pattern of the elements of secondary structure. The statistical significance of this consensus structure would be tested by the measures of reliability discussed in Section 2 3. The accuracy of the consensus structures could be tested with known structures of ATP-binding proteins.

3 4 Evolution of ATP-binding proteins. The complexity of the patterns presented suggests that ATP-binding domains sharing a pattern evolved from a common ancestor. While convergence may account for the presence of some pattern elements, it is unlikely that it could account for all of them. Therefore it appears that 41% of all sequenced ATP-binding domains evolved from one

ancestor and 31% from another. These two classes of proteins might also be related to one another since they share the Gly-x-Gly subsequence which is characteristic of many nucleotide-binding proteins. This does not mean that the entire sequences of these ATP-binding proteins evolved from common ancestors. It is more likely that ATP-binding domains are incorporated into evolving proteins as pre-fabricated modules The source of these modules could be any pre-existing ATP-binding protein. This sort of evolutionary mechanism has been discussed (Gilbert, 1978) and appears, for example, to have been involved in the evolution of blood coagulation proteins (Patthy, 1985).

While these patterns define relationships among two large groups of ATP-binding domains, they do not provide a basis for establishing degrees of relatedness between members of a group. An assessment of overall homology between pairs of sequences aligned by a pattern could perhaps establish such relationships and strengthen the argument for divergence of these sequences from common ancestors. When the sequences sharing the ATP-1 and ATP-2 patterns were compared using the FASTP program, no significant homologies were found aside from the known homologies between protein kinases. However, it remains to be seen whether the refinements in comparative methods discussed in Chapter 1 will result in the identification of additional homologies between these sequences. Moreover, restricting the search to a comparison of regions of pattern matching will increase the sensitivity of any algorithm. Significant homologies between

aligned sequences might then be detectable by pairwise comparisons.

This chapter has focused on the comparison of sequences with structural data in order to demonstrate that elements known to mediate the binding of ATP in specific proteins are generally conserved in both identity and position. The two patterns composed of these elements have been shown to correlate with the binding of ATP and probably reflect the structures of ATP-binding domains. These are the type of results expected of kinship analysis and the two methods are complementary. In the case of the ATP-2 pattern, the clustering of elements suggests that this pattern could have been derived by kinship analysis. In contrast, it seems unlikely that the scattered elements of the ATP-1 pattern could have been identified without relying on structural data. It should be possible, however, to modify the methods of kinship analysis in order to use structural data relevant to ligand binding in the comparison of sequences. In this way, an approach which has been successful for ATP-binding proteins could be rapidly extended to other ligand-binding activities.

Chapter IV

Further Analysis of Nucleotide-binding Sites: GTP-binding Proteins and Polymerases

The results obtained with ATP-binding proteins suggest that divergence of proteins from a common ancestor can be traced by comparing their sequences and accounts for the origins of many nucleotide-binding proteins. This point could be further demonstrated by extending the analysis to other nucleotide-binding proteins. Therefore analyses of GTP-binding sites and the nucleotide-binding sites of nucleic acid polymerases were done. These two types of binding sites were of interest because of their different properties. There are a number of GTP-binding proteins which are specific for GTP and a number which are somewhat less specific, tending to bind either GTP or ATP. In contrast, the nucleotide-binding sites of polymerases are necessarily non-specific, accomodating any one of the four nucleotides required for template-directed replication. Interestingly, these two different types of nucleotide-binding proteins appear to share a common evolutionary origin with one class of ATP-binding proteins.

4.1 Analysis of the sequences of GTP-binding proteins: The goal of the analysis is the derivation of a pattern that predicts which proteins bind GTP. In addition to estimating the

probability that a protein binds GTP, the pattern should have two more properties: 1) it should distinguish between those proteins which specifically bind GTP and those which bind ATP, and 2) it should indicate which proteins bind either nucleotide This goal is partially met in that a pattern for the binding of GTP is derived and proven to be valid. However, the goal of predicting nucleotide specificity cannot be met at present Two factors contribute to this shortcoming. First, there are relatively few sequences of proteins available for analysis which specifically bind GTP with the result that only a few pattern elements can be identified with confidence. The second factor is that the pattern for GTP is very similar to the ATP-2 pattern described in Chapter 2. While this similarity is of some interest and will be discussed, it also complicates the analysis.

The basic method of pattern derivation is essentially the same as that for ATP. However, an additional procedure is introduced which tests the significance of a pattern in the course of its derivation by calculating confidence levels for each pattern element. This procedure is general for the derivation of any type of consensus pattern

The set of sequence data available for this analysis included the GTP-specific proteins: elongation factor Tu (EF-Tu), β-tubulin, glutamate dehydrogenase, α-transducin (Lochrie et al., 1985), amidophosphoribosyltransferase, and the oncogene ras. The set also contained proteins which use GTP as efficiently as ATP including the dnaB gene product and the β-subunit of RNA polymerase. Structural data on GTP-binding proteins were lacking

until the three-dimensional structure of the EF-Tu•GDP complex was recently solved (la Cour et al, 1985; Jurnak, 1985). These data, along with mutational data from the oncogene ras, provided the means to compare the sequences of GTP-binding proteins.

4.1 1 An Initial Pattern for GTP-binding: The analysis began with the observation that the majority of the GTP-binding proteins contain matches to the ATP-2 pattern. Individually, these proteins tended to have only fair matches to the consensus pattern; yet when these sequences were considered as a class, their frequency of matching suggested a relationship between the ATP consensus pattern and GTP binding. In the case of EF-Tu, the pattern match corresponds well with the known structure of the GTP-binding site (la Cour et al, 1985; Jurnak, 1985), as described below.

Each element of the ATP-2 pattern was compared with the published structures of EF-Tu to ascertain whether it might be a constituent of the ligand-binding site. Two elements, b7 and b9 from Table 3.3, resided on segments of the polypeptide chain which were far from the ligand-binding site of EF-Tu and were therefore deleted from the consensus pattern. The element b7, a cationic residue at position -11 or -12, was defined by Walker et al on the basis of their original comparison of the homologies between ATP-binding proteins. It is not known to be involved in ligand-binding but might contact the phosphate groups of ATP. In F1 ATPase, the element b9 (Glu at 26 to 31) is implicated in ATP-binding by modification studies as described in Section 3.2.

While the role b9 plays in the binding of ATP is not known, one possibility is that the carboxylate anion contacts the amino group of adenine. Therefore, this element may be specific for ATP binding.

4 1.2 Generation of a ranked list of similarities· The modified pattern is shown in Table 4.1 The set of sequences of GTP-binding proteins were searched for this modified pattern and 6 of the 8 sequences were found to have significant matches to it. These sequences were: EF-Tu, β-tubulin, glutamate dehydrogenase, transducin, the oncogene ras, and the dnaB gene product. The similarities between EF-Tu, ras, and transducin and the ATP-correlated pattern described by Walker et al. (1982) have been observed previously (la Cour et al., 1985). These 6 sequences were aligned by their pattern matches, as shown in Fig. 4.1, and compared using the similarity program described in Section 3.1.1 with the results shown in Table 4.2. Note that the two highest ranking similarities have P-values which are much smaller than the rest. These small P-values imply that these similarities are important for GTP binding and this is supported by the structural data as discussed below.

4.1.3 Comparison with structural data: The most detailed information on the binding of GTP comes from the structure for the EF-Tu•GDP complex. EF-Tu rapidly hydrolyzes GTP, forming a tight complex with GDP which requires the action of elongation factor Ts for its removal. Crystals of native EF-Tu have yet to

Table 4.1: Initial Pattern for GTP-binding Proteins

Element No	Description	p	q
g1	G at 0	0.074	0.926
g2	K at 1	0.064	0.936
g3	T at 2	0.061	0.939
g4	I or V at 9	0.116	0.884
g5	G at -2	0.074	0.926
g6	G or A at -5	0.156	0.844
g7	L, I, or V at -7	0.204	0.796

This pattern is essentially the same as the ATP-2 pattern in Table 3.3, differing only in the absence of the ATP-2 elements b7 and b9. These two elements are known to be uninvolved in the binding of GDP to elongation factor Tu, as described in the text.

Figure 4.1. Alignment of GTP-binding proteins containing matches to the initial pattern

```
        #
         -10.....0.... ...10. ......20-/^/-40.....50.. ... .
        @   LI A            I
pattern      V G  G GKT     V
             | |  | |||     |
EF-Tu   HVNVTIGHVDHGKTTLTAAITTVLAKTYGGAA-/^/-INTSHVEYDTPTRHYAHVDC

c-ras   EYKLVVGAGGVGKSALTIQLIQNHFVDEYDPT-/^/-DILDTAGQEEYSAMRDQYMR

β-Tbln  QLTHLGGGTGSGMGTLLISKIREEYPDRIMNT-/^/-ATLSVHQLVENTDETYCIDN

α-Tdcn  TVKLLLGAGESGKSTIVKQMKIIHQDGYSLEE-/^/-IVRAMTTLNIQYGDSARQDD

Glu DH  RVKAIIAEGANGPTTPQADKIFLERNIMVIPD-/^/-LNHVSYGRLTFKYERDSNYH

dnaB    SDLIVAARPSMGKTTFAMNLVENAAMLQDKPV-/^/-SLSRVDQTKIRTGQLDDEDW
              *     *                                          *
        -10.....0.....10.....20-/^/-40.. .....50 .....
```

@ The initial pattern for GTP binding from Table 4.1. The alternative residues allowed for some elements are shown in clusters; for example the leftmost element can be L, I or, V. Matching residues in the GTP-binding sequences are shown in bold-faced print.

The numbering is with respect to the initial pattern. A deletion of 18 residues was made from all sequences and is marked by /^/.

* Residues of interest in the comparisons of these sequences with structural data are underlined.

Abbreviations: EF-Tu is Elongation factor Tu; β-Tbln is β-Tubulin; α-Tdcn is α-Transducin; Glu DHase is Glutamate dehydrogenase

Table 4.2: Ranked list of similarities between six GTP-binding proteins

Rank	Description of Similarity	1 Frequency	2 Yield	3 P-value
1	T at 3	0.0648	5	0.00000649
2	D at 57	0.0655	4	0.000248
3	P at -13	0.0385	3	0.00104
4	P at 19	0.0385	3	0.00104
5	K at -21 to -20	0.0999	4	0.00127
6	G or P at -3	0.1083	4	0.00168
7	K at -10	0.0513	3	0.00240
8	C at 65	0.0150	2	0.00322
9	R at 116	0.0605	3	0.00386
10	T at 50	0.0648	3	0.00469
11	D at 54	0.0655	3	0.00484
12	I at 69	0.0677	3	0.00530
13	G at -3	0.0691	3	0.00563
14	V at -11	0.0719	3	0.00631
15	H at 12	0.0214	2	0.00647

1) The expected frequencies for similarities were calculated with the binomial formula as in Table 3.1.

2) The six sequences compared were EF-Tu, β-tubulin, ras, dnaB, α-transducin, and glutamate dehydrogenase.

3) The P-value for the similarity is calculated with the binomial formula as in Table 3.1.

be grown, but two groups have solved the structure of a proteolytic fragment which contains the GDP-binding site and their results are very similar (la Cour et al., 1985; Jurnak, 1985). A schematic of the structure is shown in Fig. 4.2.a Eleven residues are cited by these authors as being directly involved in the binding of the ligand. They are listed in Fig.4.2.b along with descriptions of their roles in ligand binding.

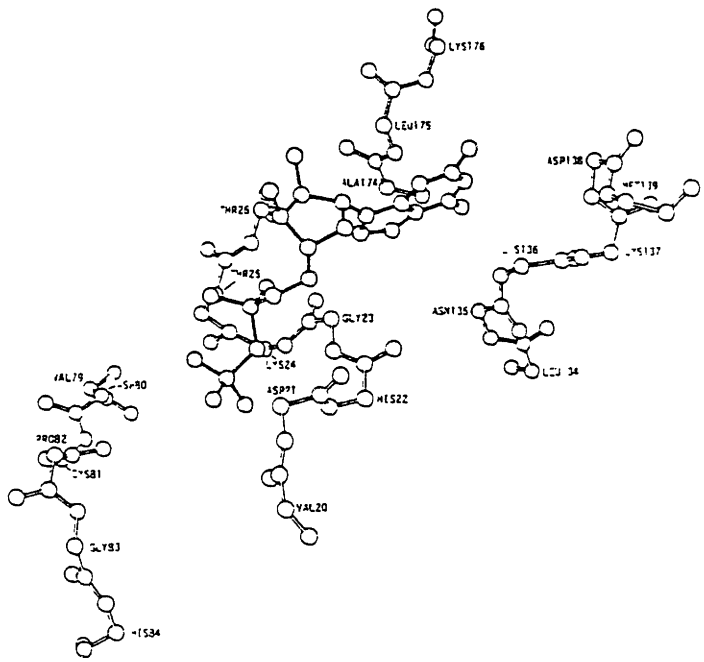
The mutational data of the oncogene ras was also examined as a possible source of structural information. While there is no direct evidence that these mutation-prone residues contact GTP, the available data support their involvement. Point mutations at position 12 are sufficient to confer a transforming phenotype (Dhar et al., 1982). Mutations at this site have been recovered repeatedly from independently arising tumors. Site-directed mutations at position 12 which convert the wild-type glycine to any other amino acid except proline result in a transformed phenotype when these proteins are assayed in tissue culture (McGrath et al., 1984). Comparison of the GTP- or GDP-binding activities of the wild-type protein with those of mutants at position 12 showed little differences between the proteins (Manne et al., 1984). However, when mutant proteins were assayed for GTPase activity, they were found to be ten-fold less active than wild-type protein (Manne et al., 1985). The conclusion was that depressed GTPase activity was associated with the transformed phenotype. While all variants at position 12 were not assayed for GTPase activity, these results imply that either glycine or

Figure 4.2: Structure of the EF-Tu•GDP Complex

a. A diagram of the EF-Tu•GDP complex (Jurnak, 1985) The residues Gly-23, Lys-24, and Thr-25 are at the core of the initial pattern in Table 4.1.

b. A description of the 11 residues cited by the authors as being involved in the binding of GTP. The positions of these residues with respect to the pattern is shown as are the designation of elements which will be added to the initial pattern to produce the final GTP pattern.

a.



b.

Residue	Pattern Elm.	Pos	Role in GDP-binding
Gly 23	g1	0	amide backbone contacts α-phosphate
Lys 24	g2	1	neutralizes the charge on one phosphate
Thr 26	g8	3	sidechain interacts with guanine
Asp 80	g9	57	carboxylate coordinates divalent cation
Ala 107		84	hydrophobic interaction with guanine
Phe 133		110	hydrophobic interaction with guanine
Leu 134		111	hydrophobic interaction with guanine
Asn 135		112	hydrophilic interaction with guanine
Lys 136		113	neutralizes the charge on one phosphate
Asp 138		115	interaction with guanine amino group
Leu 175		152	hydrophobic interaction with guanine

proline are required at position 12 for optimal GTPase activity. Another point mutation, the alteration of Gln at position 61, also confers a transformed phenotype (Taparowsky, E. et al., 1983). While this mutation has not been as well studied as the mutation at position 12, it may result in the same depression in GTPase activity.

Following the approach established with ATP-binding proteins, the sequences of the GTP-binding proteins were compared with the structural data. The most striking feature of this comparison is that the first two similarities in the ranked list of Table 4.2 correspond to residues which contact the ligand in EF-Tu. The first of these, Thr at position 3 in the pattern, is a constituent of the pocket accomodating the guanine ring The second residue, Asp at position 57 of the pattern, forms a salt bridge with the GDP-associated Mg^{2+} ion. In addition, two residues which contact the ligand correspond to the central elements, g1 and g2, of the core pattern in Table 4.1 These residues are Lys(24), g2, which neutralizes the charge of one of the phosphate groups, and Gly(23), g1, whose amide group is in close contact with the α-phosphate group of GDP. Also, the sixth similarity in the ranked list, Gly or Pro at position -3, corresponds to the only amino acids at residue 12 of the ras gene which do not activate its transforming potential. The concordance between sequence similarities and the structural data is encouraging and is supported by further analysis.

The comparison of the structural data from Fig. 4.2b with the aligned sequences is presented in a different fashion in

Table 4.3. Residues involved in the EF-Tu•GDP complex are considered first The first two residues listed correspond to the core pattern elements g1 and g2. The next residue (T at 3) contacts the guanine ring in the EF-Tu•GDP complex and 4 of the 5 sequences contain identically positioned threonine residues Subsequent analysis will justify the acceptance of this residue as a pattern element. The next residue (D at 57) contacts the GDP-bound Mg^{2+} ion and 3 of the 5 sequences contain identically placed aspartate residues. However, the remaining 2 sequences contain aspartate at position 54, suggesting that a better description of the pattern element might be "D at 54-57". The analysis below is designed to choose between these two element descriptions. The remaining ligand-contacting residues identified in EF-Tu do not recur in the aligned sequences with significant frequencies. Nonetheless, two of these will be analyzed further to demonstrate the criteria for rejecting a potential element. Of the two residues which might be involved in the binding of GTP by the oncogene ras, the similarity corresponding to the non-transforming variants of position 12 in ras will be chosen as a pattern element.

4.1 4 Derivation of a pattern for GTP-binding proteins: The data in Table 4.3 were analyzed with methods similar to those for the assessment of homology discussed in Section 1.4. Taking element g8 as an example, the question is whether observing that 4 of the 5 GTP-binding sequences contain Thr at position 3, which is analogous with the ligand-contacting Thr in EF-Tu, is

Table 4.3· Summary of the comparison of six sequences with structural data relevant to GTP binding

Ligand-contacting Residues	Element	Glu DHase	α-Tdc	dnaB	β-Tbln	ras	EF-Tu
			- Elongation factor Tu -				
G at 0	g1	+	+	+	+	+	.. ..
K at 1	g2	-	+	+	-	+	.. .
T at 3	g8	+	+	+	+	-	....
D at 57	g9	54	57	54;57	57	54	.. .
A at 84		84	-	-	85	-	...
F at 110		-	Y110	-	114	-	...
L at 111		-	+	+	-	-	.. ..
N at 112		Q113	-	Q110	N110	Q114	. ...
K at 113		-	-	-	-	-	... .
D at 115		-	113	-	-	117	.....
L at 152		-	-	153	-	153	. ...
			- ras -				
G at -3	g10	G	G	P	-	....	-
Q at 46		-	49	45	45	....	-

All sequences were aligned by their pattern matches as in Fig. 3.1 and the numbering is with respect to the pattern.Residues which directly match with the structural data are marked with (+). Residues which only appear to be similar to residues involved in ligand binding are further described by their positions and identities

significant. Reading the alignment in Fig. 4.1 vertically at position 3 results in the sequence A-T-T-T-T However, the strongest support for this element would be if all 5 sequences contained g8· T-T-T-T-T. The comparison of these two subsequences provides the test of significance. The test for significance, then. is equivalent to assessing an alignment which contains no gaps. The element g9 is a more complicated example The strongest support for the acceptance of g9 as an element would be if all 5 sequences contained an Asp at position 57: D-D-D-D-D. From Fig 4.1 it is found that 3 of the 5 sequences contain the analogous Asp, forming the sequence: M-D-D-Y-D. This observation can be assessed as above. However, the remaining 2 sequences contain Asp at position 54 and the question here is whether these residues are analogous with those at 57. In this case, reading vertically results in a subsequence which contains gaps:

```
|-D-D-|-D
|     |
|     |
D     D
```

Again, the significance of this outcome is tested by comparing the sequence above to the sequence which best supports the hypothesis, D-D-D-D-D, using the methods in Section 1.4

The results of the comparison between structural data and the aligned sequences are summarized in Table 4.4. The analysis can be described by discussing each column of the table in turn The elements are described in the first column. The values for (freq) are the probabilities of observing the element at any one

Table 4.4: Analysis of Structure-Sequence Correlates

Description	Element	freq	Yield	P-value	G	Σ-value	T-value
		- Elongation factor Tu -					
T at 3	g8	0.065	4/5	8.4×10^{-5}	1	8.4×10^{-5}	7.5×10^{-4}
D at 57	--	0.066	3/5	2.5×10^{-3}	1	2.5×10^{-3}	2.2×10^{-2}
D at 54-57	g9	0.066	5/5	1.2×10^{-6}	49	6.1×10^{-5}	5.5×10^{-4}
L at 111	--	0.088	2/5	6.5×10^{-2}	1	6.5×10^{-2}	4.5×10^{-1}
N or Q at 110-114	--	0.077	4/5	1.6×10^{-4}	375	6.0×10^{-2}	4.2×10^{-1}
		- ras -					
G or P at -3	g10	0.108	3/5	1.1×10^{-2}	1	1.1×10^{-2}	2.2×10^{-2}
Q at 45-49	--	0.040	3/5	6.0×10^{-4}	63	3.7×10^{-2}	7.3×10^{-2}

position and are just the molar frequencies for given amino acids. The number of sequences containing an element are then listed. P-values are the probability of the observed yield of sequences containing the element irrespective of position and is analogous with P(s) in Section 1.4. Σ-values are analogous with P(a) of Section 1.4, taking account of residues that occur at different positions. T-values are analogous with P(sim). In Section 1.4, P(sim) was calculated from P(a) and T, the number of trials or the number of alignments in a comparison matrix which have been assayed for similarity. In this case, the number of trials refers to the number of residues known by structural data to be involved in ligand binding.

The resulting T-values are the basis upon which elements are accepted or rejected; the critical value of T for accepting or rejecting a pattern element can be chosen by the investigator. By choosing low critical values ($T < 0.05$), the validity of pattern elements can be demonstrated in the course of pattern derivation.

The three significant pattern elements of the sequences of GTP-binding proteins are marked in Table 4.4; all three have T-values less than 0.03. The similarity "D at 54-57" was chosen as an element in preference to "D at 57" as a result of this analysis; even though the latter similarity corresponds to the specific structural feature, its probability of occurring by chance is more than 100 times that of the more broadly defined element. Note that two similarities, "L at 111" and "N or Q at 110-114", are rejected as elements, having T-values of approximately 0.5. Aside from valid elements, these were the most

promising similarities between the sequences, yet their values approximate those expected by chance. they are ultimately rejected as unreliable on the basis of their T-values. In addition, the similarity corresponding to the activating mutation at position 61 in ras is also rejected as an element because its T-value is above the critical value of 0.05. This does not mean that this residue is irrelevant to GTP-binding but that this relevance is not demonstrated by the available data. Further sequence data may support its acceptance as an element.

The three newly defined elements along with the elements of the core pattern (Table 4.1) define the GTP-correlated pattern shown in Table 4.5. The roles which these elements are expected to play in ligand binding will be discussed below. First it is worth noting the arrangement of the new elements within this pattern compared with the structural features which are not included. In Fig. 4.2, the first 4 residues of EF-Tu known to contact the ligand were accepted as pattern elements and are the closest in the EF-Tu sequence to the initial pattern elements in Table 4.1. In contrast, the remaining seven structural features were rejected as pattern elements and are all C-terminal to the accepted elements. Similarly, the element defined by the comparison with the ras protein occurs at position -3 of the pattern. It appears, then, that there is a region of conserved structural elements in these GTP-binding protein which extends from approximately -10 in the pattern to +60.

Five of the EF-Tu structural features which are omitted from the pattern reside on one loop of the primary structure and are

Table 4.5: Consensus Pattern for GTP-binding Proteins

Element No.	Description	p	q	
g1	G at 0	0.074	0.926	
g2	K at 1	0.064	0.936	
g3	T at 2	0.061	0.939	
g4	I or V at 9	0.116	0.884	
g5	G at -2	0.074	0.926	
g6	G or A at -5	0.156	0.844	
g7	L, I, or V at -7	0.204	0.796	
g8	T at 3	0.065	0.935	new elements
g9	D at 54-57	0.237	0.763	
g10	G or P at -3	0.108	0.892	

arranged in the sequence "FLNK.D". While not obvious in the schematic in Fig. 4.2.a, this loop is in contact with GDP as described in Fig. 4.2.b. Clearly significant homologies between this region and similar subsequences in elongation factor α1, ras and ras-related proteins, and transducins have been noted by a number of authors (la Cour et al., 1985; Halliday, 1983). These homologies do not occur at consistent positions relative to the GTP-correlated pattern. An effort was made to find significant matches to this subsequence in other GTP-binding proteins since the programs developed for identifying pattern matches are also useful for identifying and quantifying the significance small regions of contiguous homology. However, in a series of searches in which the stringency of matching was varied, no additional sequences of GTP-binding proteins were identified with significant matches to this subsequence. Therefore, this subsequence appears to be common to only a subset of GTP-binding proteins and cannot be included in a consensus pattern general for all GTP-binding proteins. The proteins which do contain the "FLNK.D" homology share a common property which is the tight binding of the reaction product GDP. It is possible that the subsequence is important for this similarity in the function of these proteins.

4.1.5: Correlation of the the pattern with GTP binding

Having derived a pattern for GTP-binding proteins, one would like to test its efficiency as a predictor with additional sequences not used in the pattern derivation. While these tests

were carried out, the results are not as compelling as in the case of ATP-binding proteins. A number of factors contribute to the confounding of this method of pattern verification. The first factor is the lack of sequenced proteins which are known to bind GTP specifically, only a few such proteins were found in the database in addition to those used to derive the pattern. Secondly, the binding of GTP is a function which is very similar to the binding of ATP and a number of proteins are known to bind both. It has proved to be a difficult task to catalog the data on nucleotide specificity for the proteins in the sequence database Not only are the data difficult to find in the literature, but those data are often not sufficient to answer questions of specificity.

The first step in the analysis was the identification of those proteins in the database known to bind GTP. The size of the database precluded an exhaustive search for information on every protein in it and a sampling method was used instead. Data was retrieved from the Medline Data Retrieval Service covering the years 1965-1985. Abstracts referring to guanine nucleotides were identified and those abstracts which also referred to viruses, bacteriophages, or descriptive enzyme names (ie. synthetase, synthase, dehydrogenase, etc.) were retrieved. These abstracts were then searched for data on GTP binding.

The search of the literature produced only five additional proteins with a reasonably specific affinity for guanine nucleotides. The results of searching for the GTP-correlated pattern within these test sequences is shown in Table 4.6.

Table 4.6: Correlation of the Pattern with GTP Binding

Protein	Size	S-values GTP (a)	S-values ATP (b)
Glutamate dehydrogenase[c]	503	2.8×10^{-4}	$>5.0\times10^{-4}$
RNA polymerase, β subunit	1342	2.8×10^{-4}	$>5.0\times10^{-4}$
purF, E. coli	504	-	$>5.0\times10^{-4}$
purF, B. subtilis	476	3.5×10^{-4}	$>5.0\times10^{-4}$
HGPRT, human	217	3.7×10^{-4}	2.3×10^{-4}

Total number of residues searched = 3042

a) S-values for the GTP pattern were calculated from the cumulative distribution function generated from the pattern in Table 4.5 as described in the text
b) These are S-values for matches to the ATP-2 pattern calculated as described in Section 3.2.
c) This is a second match to the GTP pattern and corresponds with a second GTP-binding site in glutamate dehydrogenase

The correlation of pattern matches with GTP binding is tested with the binomial formula. The question is whether observing 4 pattern matches with S-values of 0.00037 or less is significant given that a total of 3,042 residues were searched for the pattern:

$n = 3042$ = No. of residues searched
$m = 4$ = No. of matches observed
$p = .00037$ = S-value of worst match

$$P = \sum_{i=m}^{n} (p)^i (1-p)^{(n-i)} \binom{n}{i}$$

$$= \sum_{i=4}^{3042} (0.00037)^i (0.99963)^{(3042-i)} \binom{3042}{i}$$

$$= 0.028$$

Supporting the correlation of the pattern with GTP binding.

Glutamate dehydrogenase appears in this list because it possesses two separate GTP-specific binding sites which mediate the inhibition of the enzyme. The match listed occurs in a region of sequence different from that used in the derivation of the pattern. The purF gene product (amidophosphoribosyltransferase) also contains a GTP-specific binding site which mediates inhibition. While they share homology, the sequences from both E. coli and B. subtilis are listed because only one of the sequences contains a match. The β-subunit of RNA polymerase is included because it uses GTP as efficiently as it does ATP. The HGPRT sequence is included because it catalyzes the phosphoribosylation of guanine. Four of the five sequences are found to contain matches to the GTP pattern. The analysis of the significance of this outcome gives a P-value of 0.028, indicating a significant correlation. Nonetheless, there are some problems with this analysis. Because the patterns for GTP- and ATP-binding are so similar, the two types of binding sites could be mistaken for one another. Both glutamate dehydrogenase and the purF gene product contain separate ADP-binding sites in addition to their GTP-binding sites and the β-subunit of RNA polymerase will bind either nucleotide. Therefore it is important to compare the matches to both the GTP- and the ATP-correlated pattern. These results are shown in the last column of the table. Note that the S-values of 4 of the 5 matches to the ATP-correlated pattern are greater than 0.0005 and are, therefore, worse matches than those observed with the GTP-correlated pattern. Only in the case of HGPRT is the match to the ATP pattern better than the match to

the GTP pattern. On the whole, then, the guanine specificity of these proteins is reflected in their pattern matches. Therefore, while less compelling than the proofs of the two ATP patterns, these results support the correlation between GTP-binding and the proposed GTP pattern.

The validity of the GTP-correlated pattern has been demonstrated both in the course of its derivation and by a subsequent test with additional sequences of proteins which specifically bind GTP. However, a problem remains with its use as a predictor and that is the issue of nucleotide specificity. To assess the ability of the pattern to distinguish nucleotide specificities, the 1985 Protein Sequence Database was searched for matches to the GTP pattern. Proteins with matches were ranked in order of the quality of those matches. The highest ranking matches are shown in Table 4.7. The table also shows which of these proteins are known to bind ATP or GTP as well as which proteins are GTP-specific. It is clear that the pattern for GTP does not distinguish GTP-specific proteins from proteins which bind ATP. This is not surprising since the pattern for GTP is so similar to the second ATP pattern.

While the similarity of these two patterns complicates the prediction of nucleotide-binding activity, it is interesting from an evolutionary point of view. The results of searching the 1985 database support a conjecture that proteins with matches to the ATP-2 pattern, and not to ATP-1, tend to display a broad range of nucleotide affinities. Therefore, the ATP-2 pattern might reflect a general nucleotide-binding site which can be modified to confer

Table 4.7: Ranked list of proteins with GTP pattern matches compared with nucleotide affinities

Protein	S-value	Affinity ATP	Affinity GTP
Thymidine kinase, Herpes	2.4E-09	+	-
dnaB protein - Escherichia coli	4.7E-09	+	+
Adenylate kinase	9.9E-09	+	+
183.3K protein, Tobacco mos. virus	9.2E-08	+	+
Nitrogenase, iron protein	2.2E-07	+	?
Glutamate dehydrogenase	3.6E-07	-	+
ras protein	5.6E-07	-	+
recA protein - Escherichia coli	1.1E-06	+	-
uvrD protein - Escherichia coli	1.5E-06	+	?
Myosin heavy chain - Nematode	1.5E-06	+	+
F1 ATPase, beta chain	1.5E-06	+	+
recF protein - Escherichia coli	3.3E-06	+	?
Elongation factor 1 alpha	5.8E-06	-	+
Tubulin beta chain	7.3E-06	-	+
Glycyl-tRNA synthetase	1.0E-05	+	?
Elongation factors Tu	1.3E-05	-	+
Elongation factor G	1.5E-05	-	+
Histidine permease inner membrane prot.	1.8E-05	+	?
dnaA protein - Escherichia coli	1.9E-05	+	?
Transcription termination factor, rho	3.6E-05	+	+

The righthand columns characterize nucleotide affinities, + means the protein has been shown to bind the specified ligand, - that it has been shown not bind the ligand, and ? that no information is either available or retrievable.

specificity. The ATP-1 pattern, on the other hand, might reflect a more specialized nucleotide-binding site which is more specific for ATP.

The basic observation is that those nucleotide-dependent proteins with a demonstrated affinity for GTP tend to contain matches both to the GTP pattern and to the ATP-2 pattern but not to the ATP-1 pattern. The list of GTP-binding proteins includes those proteins which specifically bind GTP along with those which use either GTP or ATP. The list was compiled from a computerized search of the literature using the sampling method described above. A total of 20 sequenced proteins which are known to bind GTP were produced in this search. Of these 20 proteins, 11 contain matches to either the GTP pattern or the ATP-2 pattern. In contrast, only 1 of the proteins, adenylate kinase, also contains a match to the first ATP pattern. The significance of this observation was tested with the hypergeometric formula. The resulting P-value of 0.01 suggests a strong tendency for GTP-binding proteins to contain matches only to variants of the ATP-2 pattern.

Despite this test, the observation is still subject to some reservations and therefore only reflects a suggestive trend in the data. At issue is the classification of proteins according to their affinity for GTP. The literature search distinguished proteins cited as having an affinity for GTP from those not cited. A good test of the hypothesis would require data on nucleotide affinity for all nucleotide-dependent proteins in the database. Often however, the available data only measures the

relative efficiency of various nucleotides as cofactors in a given enzymatic reaction. Assaying each nucleotide in turn, the Michaelis constants Km and Vmax are used to characterize the nucleotide specificity of an enzyme. Yet assuming that ATP is the preferred nucleotide for a given enzyme, another nucleotide might be a less efficient cofactor for a number of reasons, including: 1) the nucleotide actually binds less tightly, or 2) the enzyme has a greater affinity for the product NDP than it does the substrate NTP. Measuring the inhibition of an ATP-dependent reaction by other nucleotides would be one way of determining the affinity of an enzyme for various nucleotides, ligand-binding assays are another. However, this sort of affinity data does not appear to be available for most enzymes. An exhaustive search of the literature would probably yield nucleotide specificity data for most enzymes. The particular question to be answered is whether proteins with the ATP-1 pattern have a marked preference for ATP as their cofactor.

4.2 An analysis of the nucleotide-binding site of polymerases:

The hypothesis that the ATP-2/GTP pattern reflects a generalized nucleotide-binding domain was further tested by comparing the sequences of nucleic acid polymerases. An essential feature of the ligand-binding sites of polymerases is that they can accomodate any of four nucleotide triphosphates. It can be shown that the nucleotide-binding sites of a number of polymerases contain a variant of the GTP and ATP-2 patterns supporting the hypothesis that these patterns are elaborations of a generalized

nucleotide-binding site. None of the published polymerase sequences contain especially significant matches to either established pattern. Yet for one protein, DNA polymerase I, the pattern match corresponds to that portion of the sequence known, by crystallographic analysis, to constitute the nucleotide-binding site. Moreover, when the sequences of other polymerases are aligned by these matches an interesting pattern of homology is found which corresponds with those residues known to contact the ligand of the crystallized enzyme.

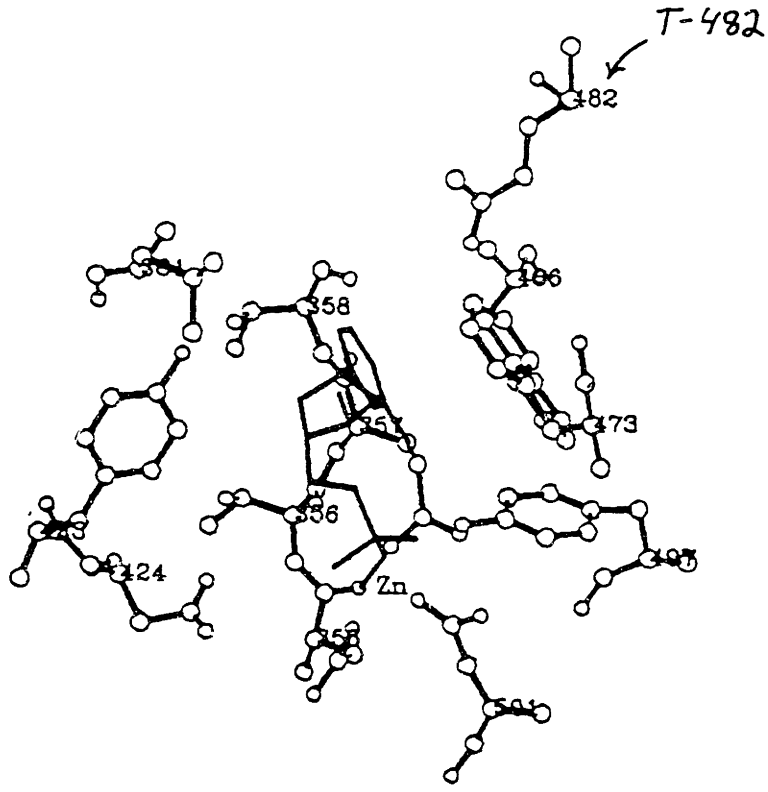
The structure of the Klenow fragment of DNA polymerase I complexed with dTMP has recently been solved to 4 3-Å resolution (Ollis et al., 1985a). The fragment is composed of two domains. The 200 N-terminal residues form the nucleotide-binding domain and the remaining 400 C-terminal residues form the DNA-binding domain. While the resolution is not sufficient to identify most nucleotide-enzyme interactions, the interactions of five residues with the ligand were noted by the authors. In addition, the interactions of three anionic residues with a metal ion were described. This enzyme-bound metal ion is also coordinated by the α-phosphate of dTMP and is essential for catalytic activity. The metal-binding site will accomodate a number of divalent cations such as Mn^{+2} and $Mg^{+2}$and is not specific for Zn^{+2} as was previously reported. A diagram of the binding site is shown in Figure 4.4. The residues cited by the authors as interacting with the ligand are also listed along with descriptions of their proposed functions.

Figure 4.3: Structure of the DNA Polymerase I dTMP-Binding Site

a. A diagram of the nucleotide-binding site of the Klenow fragment of DNA polI. Only a few residues are shown and one of them, Thr-482, is element 3 of a match in the sequence of polI to both the ATP-2 and GTP patterns.

b. Descriptions of the 8 residues cited by the authors as being involved in the binding of dTMP or in the binding of an essential, enzyme-bound divalent cation. Numbering with respect to nucleotide patterns is shown in parentheses.

a.



b.

Residue		Description
1) Asp 355	(-125)	carboxylate coordinates the divalent cation
2) Glu 357	(-123)	carboxylate coordinates the divalent cation
3) Thr 358	(-122)	backbone amide contacts 3'-OH of dTMP
4) Leu 361	(-119)	side chain contacts the dTMP base
5) Asp 424	(-56)	interacts with α-phosphate through a small molecule
6) Phe 473	(-7)	side chain contacts second side of base
7) Tyr 497	(17)	hydroxyl interacts with the α-phosphate
8) Asp 501	(21)	carboxylate coordinates the divalent cation

The sequence of polI was searched for the core pattern in Table 4.1. Since this pattern is composed only of the elements which the ATP-2 and GTP patterns have in common, it represented the best approximation of a general pattern for nucleotide-binding sites. The best match to this pattern occurs at position 480 in the polI sequence, within the nucleotide-binding domain, the S-value for this match is 0.0002. Given that the nucleotide-binding domain is composed of 200 residues, the significance of this match is 0.04, supporting the relevance of this match to nucleotide binding. The relevance of the pattern match to the structure of the binding site is further supported by the crystallographic data. While interactions between pattern elements and dTMP are not apparent in this 3.3Å structure, the segment of protein containing the match is in proximity to the nucleotide as shown in Figure 4.3. Therefore, there is good agreement between the location of the nucleotide-binding site predicted from the sequence with that observed in the crystal structure.

The authors of the crystallographic study also compared the sequence of polI with that of the DNA polymerase of phage T7 (Ollis et al., 1985b). They describe 9 regions of homology between these sequences, all but one of which occur in the large DNA-binding domain. The alignments of these sequences were presented without the statistical data required to judge their significance. Nonetheless, a number of these regions are homologous by the criteria of Section 1.4; these regions lie within the DNA-binding domain. No significant homology was found

in the nucleotide-binding domain. However, I was able to find significant similarities by focussing on the results of the crystallographic analysis. As before, the sequence of T7 pol was searched for matches to the pattern in Figure 4.4. The best match was observed at position 116, with an S-value of 0.0008. In itself, this is not a very significant match yet it serves to align the two sequences giving the results shown in Table 4.8. The point of alignment is indicated by the pattern above the sequences. Residues in polI known to contact either the nucleotide or the essential divalent cation are indicated with asterisks. In this alignment, 5 of the 8 residues in contact with the ligands of polI are conserved in T7 pol. Furthermore, the tyrosine residue in the subsequence "KYD", while not cited as a contact residue, is clearly within the binding cleft as shown in Figure 4.4. The gaps required for aligning each region of homology were: 13 for region a, 7 for region b, and 0 for region c.

The significance of the relationship between the sequences was assessed with the methods in Section 1.4. In this case the alignment between the sequences is determined by their respective pattern matches so that P(a) is equivalent to P(sim) and is the value of interest. The most conservative estimate of significance ignores the Tyr residue not specifically cited by the authors. In this case: 5 of the 8 polI residues involved in ligand binding are conserved in T7 pol and $P(s)=3.17\times10^{-5}$, G=405, and P(a)=P(sim)=0.013. However, Fig. 4.3a clearly shows that Tyr-423 of region b is in the binding cleft and this residue is also

Table 4.8: Homologies between the nucleotide-binding sites of polymerases

a. Homologies between the NTP-binding sites of polI and T7 pol

```
DNA pol I
   a          b                    c
D.ET..L..(-125)..K.G..LKYD...(-56)...FeeiaGkGKnqltfnqi-ALEeAgrYaa
* **  *              *           *                          *
D.E...L..(-112).......KYD...(-49)...dmGllrsGKlpgkrfgshALE-AwgYrl
                                    |  | |||
T7 DNA pol                          G  G GKT
```

b. Homologies between the NTP-binding sites of other polymerases

```
                    b              c
Terminal trans.  KKGLLLYYD...(-46)...sksnqqeGKTwkairvdlvmc---PYEn
                  ||||
Epstein-Barr pol KGLLKKPD...(-45)...vvGgdedGKgmwwrqraqegtarpead-
                                                 ||| |||  |
Oligo-A synthetase    KYYD...(-34)...ptGnlGgGdpkgwrqlaqeaeawlnYp

Brome mosaic pol ......KYD...(-57)...vdGvaGcGKTtaikdafrmge-------

Ribo. reductase  ......KYD...(-37)...aavkqleGKylvqnrvtge-----iYEs
                                    |  | |||
                                    G  G GKT
```

All sequences were selected by the initial GTP pattern in Table 4.1 and aligned by their pattern matches. The core of that pattern is shown below the sequences and residues matching that pattern are in boldface print.

* denotes residues in polI known to be involved in the binding of either dTMP, in the cases of regions b and c, or the essential enzyme-bound divalent cation, in the case of region c. The homology between polI and T7 DNA polymerase is the strongest, involving all three regions. Region c is missing in the remaining sequences. Additional homologies are indicated by (|).

conserved in T7 pol. Accepting that 9 residues known to be involved in ligand binding and that 6 of these are conserved yields a value of 0.0011.

The analysis indicates a clear homology between the structures of the nucleotide-binding sites of the two enzymes. This homology is composed of scattered elements and is not embedded in a larger pattern of detectable homology. As a result, the significance of this homology could only be demonstrated in a predetermined alignment of the two sequences such as that provided by the matches to the pattern in Table 4.1. Moreover, these pattern matches are relatively poor and are not obviously related to the homology between nucleotide-binding proteins described by Walker et al. (1982). A more sensitive measure of relatedness, the S-value, was required to produce this alignment.

The analysis also identifies interesting targets for site-directed mutagenesis and predicts the behavior of the mutants to some extent. In the case of the EFTu•GDP complex, the Lys at position 2 of the GTP pattern contacts the β-phosphate and a similar function for the analogous Lys in either polI or T7 pol could be postulated. Mutagenesis of this residue should decrease either enzyme's affinity for nucleotide. Moreover, a greater decrease is expected in the affinity for the triphosphate as opposed to the monophosphate nucleotide. Mutation of the anionic residues in T7 pol which are analogous to those known to coordinate the enzyme-bound divalent cation should also have a fairly specific effect. The affinity of either enzyme for divalent cations should be sharply reduced. More difficult to

predict is the effect of these mutations on nucleotide affinity. However, the effect of analogous mutations on nucleotide affinity are expected to be similar for both proteins.

This analysis was extended to other nucleic acid polymerases. As shown in Table 4.8, this attempt is partially successful. The most interesting similarities between the aligned sequences occur in region b. Four additional polymerases are homologous with polI in this region, they are: terminal transferase, the oligo-A synthetase induced by interferon, and the polymerases of Epstein-Barr and brome mosaic viruses. The homologs of the brome mosaic virus polymerase which include the tobacco, tomato, and alfalfa mosaic viruses, also share this similarity. In addition, the large subunit of ribonucleotide reductase shares this similarity with the polymerases. This is consistent with the requirement that reductase bind and reduce all four ribonucleotides. A continuation of this analysis is in progress. However, at this point, neither RNA polymerases nor the polymerases of either the retroviruses nor the picornaviruses conform to this pattern. The absence of the pattern in these RNA polymerases does not suggest that the pattern is specific to DNA polymerases since mosaic virus polymerases are also RNA-specific.

4.3 Evolution of nucleotide-binding proteins: Some features of the evolution of nucleotide-binding proteins are suggested by the comparison of their sequences. The analysis in Chapter 3 resulted in the derivation of two patterns,ATP-1 and ATP-2, that select two large classes of ATP-binding proteins. The complexity of the

patterns argues for the divergence of all members of a class from some common precursor. The patterns share two common elements which form the subsequence Gly-x-Gly. This subsequence is also shared by a number of NAD-dependent enzymes, lies within an area of structural homology between both NAD- and ATP-binding proteins and is involved in ligand binding (Rossman et al., 1977). These facts support the idea that the two classes of ATP-binding proteins are themselves related. The two principal results of this chapter are:

1) that elements of the ATP-2 pattern are also shared by GTP-binding domains and the NTP-binding domains of polymerases

and 2) that the patterns for ATP-, GTP-, and NTP-binding domains contain elements specific to their respective functions in addition to those they have in common.

The elements which these patterns share form the subsequence "G..G.GK" and will be called the C, or common, pattern.

The C pattern appears to represent a generalized nucleotide-binding domain which has been elaborated in a number of different ways to serve different functions. In contrast, the ATP-1 pattern may represent a more specific nucleotide-binding domain. The fact that the initial elements of the ATP-1 pattern were identified by comparisons of ATP-binding sequences with lactate dehydrogenase, which is in turn structurally homologous with other dehydrogenases (Rossman et al., 1977), suggests that ATP-binding domains sharing this pattern are related to NAD-binding domains. Yet this relationship supports the idea of the specificity of the

pattern since NAD contains an ADP moiety.

DISCUSSION

The methods I have described can be used to make a systematic analysis of available sequence data. A primary aim of this analysis and the focus of this discussion is the prediction of a protein's function from its sequence. The various steps of the approach will be outlined here. First, however, I will briefly summarize some salient points of the methods discussed above.

The first method is the comparison of pairs of sequences for homology. This method is based on a statistical measure of the significance of an alignment between two sequences. A program was described which will explicitly search for this most significant alignment. If this measure is below some critical value one can immediately conclude that the two sequences are homologous and arose from a common ancestor. In principle, these comparisons could be made according to any rules of correspondance but two rules are of interest here. The first accepts only identically matching residues as a basis for alignment; the second will also accept conservative substitutions.

This basic method can be extended to identify kinship between groups of sequences sharing a common function. Within kinship groups, values for the measure of similarity between sequences are, as a group, significantly lower than those expected for unrelated sequences. However, no pair of sequences need be significantly homologous for kinship to be defined over

the group. In an analogous fashion, kinship can be established between groups of functionally similar proteins. In this way, it might be possible to trace the evolution of one function from another.

When the alignments between sequences of a kinship group are examined, one expects to find a region of homology common to a majority of the sequences. This common homology could be expressed as a consensus pattern using the methods in Chapters 3 and 4. The significance of pattern matches can be assessed in a straightforward way. Structural data can also be used in the derivation of consensus patterns.

The first step in the comprehensive program is the comparison of all sequences in the database using identical matching as the basis of correspondance. All alignments with scores below some critical value are identified as homologous. Because this method of comparison is an improvement over existing methods, new homologies between sequences are likely to be discovered and previous assertions of homology will be tested. However, this step serves two additional purposes by providing a set of sequences for the analysis of molecular evolution and of conservative substitution.

The statistical measure of similarity serves equally well as a measure of divergence for an analysis of molecular evolution The sequences of cytochrome c, insulin, and ribonuclease have been determined for a large number of species. In principle, one should be able to construct a phylogenetic tree from these data which traces the course of organismic evolution. However a

central assumption of this enterprise, the neutral theory of mutation, needs testing. The hypothesis is that mutation rates for each type of protein are constant over a geological time span, although these rates might differ between proteins of different types (Kimura, 1975). Three phylogenetic trees can be constructed from the data for each of the three sets of protein sequences. If the neutral hypothesis is correct, these three trees should be virtually indistinguishable.

The collection of homologous sequences is also useful for a thorough analysis of conservative substitutions. A large and representative sample can be assembled from this set of homologs which varies in the degree of divergence between sequence pairs. By applying the methods of Sections 1.2 and 2.3 to this sample, one should be able to establish criteria of conservative substitution and determine whether these criteria apply to all proteins equally. These new criteria notwithstanding, the remainder of the program is carried out according to the criterion of identical matching. Then the program is repeated using the criteria of conservative substitution. When the results of the two programs are compared, one should see improved sensitivity with the second program but its results should not contradict the results of the first.

The second step in the program is the kinship analysis of sequences. After separating the sequences into groups sharing a common function, the methods of Section 1.5 are applied. The results with nucleotide-binding proteins demonstate that even distant relationships between sequences sharing a common function

can be identified by sequence comparison. Consequently, kinship analysis is also expected to identify many distant evolutionary relationships which are not apparent in simple pairwise comparisons. This expected utility is only partially dependent on the assumption of the constancy of mutation rates: one need only assume continuity of divergence for this approach to be useful. In many cases, examination of the sequence alignments of kinship groups should lead to the definition of consensus patterns characteristic of particular functions. The sensitivity of these analyses could be further increased by considering structural as well as sequence data, as was done in Chapters 3 and 4. In this way, a lexicon of patterns predictive of specific functions could be rapidly compiled.

Kinship groups and their resulting patterns can themselves be the subject of taxonomic analysis. While this task promises to be complicated, the results could be interesting. One would be constructing a phylogenetic tree of protein structure and, by extension, protein function. In contrast with organismic evolution, however, one does not expect this tree to be simple and it may not look like a tree at all. The results of Chapters 3 and 4 suggest that nucleotide-binding domains have been transposed throughout genomes. Similar mobile domains have been identified in blood coagulation proteins (Patthy, 1985). Therefore, protein evolution is likely to involve a large scale reassortment of mobile, pre-formed domains which recombine to form new proteins.

An interesting possibility is that the most successful bioengineering strategies will be those which try to mimic this reassortment of domains.

The original question of how generally useful will sequence comparison be for the prediction of protein function cannot be fully answered yet. However, these preliminary results suggest that it should be very useful. This is true even if the methods are especially useful only for proteins which have evolved recently. Some of the most interesting proteins, such as those involved in development and neural function, will have evolved relatively recently and should not have diverged much from their precursors. Therefore, comparative methods should be very successful when applied to these proteins. Yet the results with nucleotide-binding proteins demonstrate that useful analyses can be done even with sequences that are distantly related.

I have outlined a program for the comparison of protein sequences which refines and extends previous methods and is itself subject to improvement. As these methods evolve, they should play an increasingly prominent role in molecular biology. A practical use and test of these comparative methods is the prediction of protein function. However, the ability to trace protein phylogeny serves another function by providing a system of classification for protein sequences. In the same way that classical taxonomy confers order on the diversity observed at the organismic level, so too will taxonomic methods confer order at the molecular level.

Literature Cited

Anfinsen, C. (1967). Harvey Lecture 61, 95-116.

Aubert, J. et al. (1976). Arch. Biochem. Biophys. 175, 410-418.

Barker, W. & Dayhoff, M. (1982). Proc. natn. Acad. Sci. U.S.A. 79, 2836- 2840.

Barker,W. et al. (1984). Protein Sequence Database (Natn. Biomed. Res. Found., Washington D.C.)

Barker,W. et al. (1983). Protein Sequence Database (Natn. Biomed. Res. Found., Washington D.C.)

Beck, E. et al. (1978). Nucleic Acids Res. 5, 4495-4503.

Chothia, C. & Lesk, A. M. (1985). J. Mol. Biol. 182, 151-158.

Chou, P. & Fasman, G. (1978). Ann. Rev. Biochem. 47, 251-292.

Clertant, P. et al. (1984). J. Biol Chem. 259, 15196-15203.

Cohn, M, Leigh, J., Jr. & Reed G. (1972). Cold Spring Harbor Symp. Quant. Biol. 36, 533-540.

Cornelissen, B. et al. (1983). Nucleic Acids Res. 11, 1253-1265.

Dayhoff, M., Hunt, L. & Hurst-Calderone, S. (1978). Atlas of Protein Sequence Structure; Vol. 5, Suppl. 3: 363-369.

Dayhoff, M., Orcutt, B., & Schwartz, R. (1978b). Atlas of Protein Sequence and Structure, vol.5, suppl.3; (National Biomedical Res. Found., Washington)

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). Atlas of Protein Sequence and Structure, vol.5, suppl.3; (National Biomedical Res. Found., Washington)

Dhar, R. et al. (1982). Science 217, 934-937.

Doolittle, R, et al. (1983). Science 221, 275-277.

Doolittle, R. (1981). Science 214, 149-159.

Downward, J. et al. (1984). Nature 307, 521-527.

Egyed, A. (1975). Biochim. Biophys. Acta 411, 349-356.

Erickson, B. & Sellers, P. (1983). in Time Warps, String Edits, and Macromolecules (Addison & Wesley, Reading, Mass.) pg. 53-91; Sankoff & Kruskal, edit.

Eventoff, W. et al. (1977). Proc. natn. Acad. Sci. U.S.A. 74, 2677-2681.

Finch, P. & Emmerson, P. (1984). Nucleic Acids Res. 12, 5789-5799.

Fry, D., Kuby, S. & Mildvan, A. (1985). Biochemistry 24, 4680-4694.

Gilbert, W. (1978). Nature 271, 501-503.

Goelet, P. et al. (1982). Proc. natn. Acad. Sci. USA 79, 5818-5822.

Gram, H. & Rueger, W. (1985). EMBO J. 4, 257-264.

Grau, U., Trommer, W. & Rossmann, M. (1981). J. Mol. Biol. 151, 289-307.

Gritz, L. & Davies, J. (1983). Gene 25, 179-188.

Halliday, K. (1983). J. Cyclic Nucleotide Protein Phosphor. Res. 9, 435-448.

Hansch, C. & Leo, A. (1979). "Substitute Constants for Correlation Analysis in Chemistry and Biology." Wiley, New York.

Higgins, C., Hiles, I., Whalley, K., & Jameison, D. (1985). EMBO J. 4, 1033-1040.

Hildebrand, J. (1979). Proc. natn. Acad. Sci. 76, 194-198.

Hvidt, A. (1983). Ann. Rev. Biophys. Bioeng. 12, 1-20

Jurnak, F. (1985). Science 230, 32-36.

Khananshvili, D. & Gromet-Elhanan, Z. (1985). Proc. natn. Acad. Sci. U.S.A. 82, 1886-1890.

Kloetzer, W., Maxwell, S., & Arlinghaus R. (1983). Proc. natn. Acad. Sci. U.S.A. 80, 412-416.

Kornberg, A. (1980). "DNA Replication" (W. H. Freeman, San Fransisco).

La Cour, T., Nyborg, J., Thirup, S., & Clark, B. (1985). EMBO J. 4, 2385-2388.

Lesk, A. M. & Chothia, C. (1980). J. Mol. Biol. 136, 225-270.

Lesk, A. M. & Chothia, C. (1982). J. Mol. Biol. 160, 325-342.

Lipman, D. & Pearson, W. (1985). Science 227, 1435-1441.

Lipman, D., Wilbur, W., Smith, T. & Waterman, M. (1984). Nucleic Acids Res. 12, 215-226.

Lochrie, M., Hurley, J., & Simon, M. (1985). Science 228, 96-99

Maizel, J. & Lenk, R. (1981). Proc. natn. Acad. Sci. U. S. A. 78, 7665-7669

Manne, V., Bekesi, E., & Kung, H. (1985). Proc. natn. Acad Sci. U. S. A. 82, 376-380.

Manne, V., Yamazaki, S., & Kung, H. (1984). Proc. natn. Acad. Sci. U. S. A. 81, 6953-6957.

Matthews, B. (1975). Biochim. Biophys. Acta 405, 442-451.

McGrath, J., Capon, D., Goeddel, D. & Levinson, A. (1984). Nature 310, 644-649

McLachlan, A. & Boswell, D. (1985). J. Mol. Biol. 185, 39-49.

McLachlin, A. (1971). J. Mol. Biol. 61, 409-424.

Miles, J. & Guest, J. (1984). Nucleic Acids Res. 12, 3631-3642.

Mortenson, L. & Thorneley, R. (1979). Ann. Rev. Biochem. 48, 387-418.

Needleman, S. & Wunsch, C. (1970). J. Mol. Biol. 48, 443-453.

Nemethy, G. et al. (1981). Macromolecules 14, 975-985.

Neurath, H. (1984). Science 224, 350-357.

Ohlsson, I., Nordstrom, B. & Branden, C. (1974). J. Mol. Biol. 89, 339-354.

Ollis, D., Brick, P., Hamlin, R., Xuong, N. & Steitz, T. (1985a). Nature 313, 762-766.

Ollis, D., Kline, C., & Steitz, T. (1985b). Nature 313, 818-821.

Pai, E. et al. (1977). J. Mol. Biol. 114, 37-45.

Pal, P. & Colman, R. (1979). Biochemistry 18, 838-845.

Palau, J. et al. (1982). Int. J. Pept. Protein Res. 19, 394-401.

Patthy, L. (1985). Cell 41, 657-663.

Petersen, T. et al. (1983). Proc. natn. Acad. Sci. U. S. A. 80, 137-141.

Robson, B. & Suzuki, E. (1976). J. Mol. Biol. 107, 327-356.

Rose, G., Gierasch, L., & Smith, J. (1985). Adv. Protein Chem. 37, 1-109.

Rossmann, M. & Argos, P. (1977). J. Mol. Biol. 109, 99-129.

Rossmann, M., Moras, D. & Olsen, K. (1974). Nature 250, 194-199.

Sanger, F. et al. (1982). J. Mol. Biol. 162, 729-773.

Sankoff, D. & Kruskal, J. (1983). "Time Warps, String Edits, and Macromolecules" (Addison & Wesley; Reading, Mass.).

Schultz, G. et al. (1974). Nature 250, 140-142.

Schulz, G. & Schirmer, R. (1974). Nature 250, 142-144.

Smith, G. & Griffin, J. (1978). Science 199, 1214-1216.

Smith, T., Waterman, M. & Burks C. (1985). Nucleic Acids Res. 13, 645-656.

Sokal, R. & Rohlf, F. (1969). "Biometry", pg. 71-95 (W. H. Freeman,San Francisco).

Staden, R. (1982). Nucleic Acids Res. 10, 2951-2961.

Stura, E. et al. (1983). J. Mol. Biol. 170, 529-565.

Tamkun, J., Schwarzbauer, J. & Hynes, R. (1984). Proc. natn. Acad. Sci. U. S. A. 81, 5140-5144.

Tanford, C. (1980). "The Hydrophobic Effect", 2nd ed. (Wiley, New York).

Taparowsky, E., Shimizu, M., Goldfarb, M. & Wigler, M. (1983) Cell 34, 581-586.

Tso, J. et al. (1982). J. Biol. Chem. 257, 3525-3531.

Walker, J., Saraste, M., Runswick, M. & Gay, N. (1982). EMBO J. 8, 945-951.

Waterman, M. S. (1984). Bulletin Math. Biol. 46, 473-500.

Waterman, M., Smith, T., & Beter, W. (1976). Adv. Math. 20, 367-387.

Wolfenden, R. et al. (1981). Biochemistry 20, 849-855.

Yoshida, M., Allison, W., & Esch, F. (1981). Fed. Proc. 40, 1734-1740.

Zoller, M., Nelson, N. & Taylor, S. (1981). J. Biol. Chem. 256, 10837-10842.

Appendix

PatternSearch is a program written in ANSI Pascal which is designed to search the Natn. Biomedical Res. Found. Protein Sequence Database for matches to a consensus pattern. The sequence database must be in the original format provided by the NBRF. Two files are produced by the program. The first is a list of the proteins containing pattern matches similar to that shown in Table 4.7. This list is ranked by the quality of matching as measured by the S-value. The second file contains the relevant sequence segments of these matching proteins. The segments are aligned by the first element of the consensus pattern.

The pattern is obtained from an external file which follows a specific format. A sample file specifying the GTP pattern of Table 4.5 is shown below:

```
 < comments, if any >
$
?  1   0 [G] 0
?  2   1 [K] 1
?  3   2 [T] 2
?  4   9 [IV] 9
?  5  -2 [G] -2
?  6  -5 [AG] -5
?  7  -7 [LIV] -7
?  8   3 [T] 3
?  9  54 [D] 57
? 10  -3 [GP] -3
!  1.0E-04
#  10
:  65
```

Any descriptive comments about a pattern can be made in the lines preceeding the $ sign. Various punctuation marks are used to organize the description of the pattern:

$ signals the start of pattern specification;

? marks lines specifying pattern elements;

! specifies the significance cutoff, only pattern matches with S-values below this cutoff are saved;

specifies the number of elements in the pattern;

: specifies the point of alignment of sequence output, sequences are printed in lines of 130 characters, a value of 65 places the first pattern element in the center of the line;

each mark must be the first character of a line. Each pattern element is identified by the first number in the line and is specified by a set of amino acids to be searched for within a range of positions from the first element. For example, element 9 is Asp, ie. D, 54 to 57 residues from element 1.

Program outline: After initializing some variables, the program reads the pattern from an external file with the procedure GetElements. The cumulative probability distribution for pattern matches is then constructed by brute force. A probability value for each possible pattern match is calculated by the procedure GenerateCombinations and both the match and the value are stored in a record. A list of these records is constructed which is ranked by the associated probability values. Each pattern match

is represented by an integer value in these records. The ranked list is summed to give the cumulative distribution. The list is then sorted with respect to the integer representation of matches into a tree structure.

Searches of the database are initiated with the procedure SearchDayhoff and each database entry is entered into an array by the procedure NextSequence. Sequence arrays are scanned for pattern matches. The S-value of each match is obtained from the tree structure above. If the S-value of the match is below the cutoff, the region of matching is printed in the file of sequence output and the match is added to the ranked list of observed matches. This list is printed after the search of the database is completed.

```
PROGRAM PatternSearch
(input, output, DayIn, DayOut, RankFile, Pattern);
{*****************************************************
 * DayIn = input file: NBRF Protein Sequence Database *
 * Dayout = output file of sequences                  *
 * RankFile = output file of ranked pattern matches   *
 * Pattern = input file specifying pattern            *
 *****************************************************}


 TYPE   SubInt= -32000..32000;
        Components= SET OF 'A'..'Z';

        Frequencies= ARRAY ['A'..'Z'] OF real;
          {stores values for the average molar composition
           of proteins}

        Elements = RECORD
                     s1, s2: Components;
                     n1, c1, n2, c2: SubInt;
                     f: real;
                   END;
        ElementArray = ARRAY [1..20] OF Elements;
          {the pattern is represented by this array of records}

        SequenceArray= ARRAY[-260..6000] OF char;
        MatchString =  ARRAY [1..20] OF char;
          {matches are represented in this array by a combination
           of 1's and 0's}

        RankPointer= ^Rankings;
        Rankings= RECORD
                    Tag: integer;  { represents a MatchString }
                    S: real;       { probability value }
                    More, Less: RankPointer
                  END;
          {records for calculating the cumulative distribution
           function of pattern matches expected for random
           sequences}

        NameString= ARRAY[1..50] OF char;
        DataPointer= ^Data;
        Data= RECORD
                  Name: NameString;
                  Total, Pos: SubInt;   {Total = sequence length}
                  Tag: integer;         {Pos = match position}
                  S: real;
                  More, Less: DataPointer
                END;
          {data structure for compiling the ranked list of
           observed pattern matches }
```

```
      DataPad= RECORD
                Name: NameString;
                Total: SubInt;
                Tag, Pos: integer;
                S: real;
               END;
        {scratchpad for collecting data from sequences}


VAR  Threshold: real; {significance cutoff}
     Last,            {No. of pattern elements}
     Dx,              {point of alignment for sequence output}
     Length,          {No. of residues spanned by pattern}
     i: SubInt;
     JustPvalues: boolean;
     Elm: ElementArray;
     Freq: Frequencies;
     Seq: SequenceArray;
     Sort: MatchString;
     Cursor, Head, Tail, Root: RankPointer;
     Best, Worst: DataPointer;
     Datum0, Datum00: DataPad;
     DayIn, DayOut, RankFile, Pattern: text;


FUNCTION Pwr( x: real; a: SubInt): real;
BEGIN    Pwr:= exp(a * ln(x))  END;
```

```
PROCEDURE GetElements;

{************************************************************
* Takes the specifications of a consensus pattern from an  *
* external file and converts them to the program's format. *
* an array of records corresponding with pattern elements. *
************************************************************}

VAR        a: char;          Capitals: Components;
     i, j, k: integer;         simple: boolean;
        x, y: real;

BEGIN
  Reset (Pattern);
  Capitals:= ['A'..'Z'];
  JustPvalues:= false;
  REPEAT
    Readln (Pattern, a);
  UNTIL (a = '$');
  WHILE NOT eof(Pattern)
  DO BEGIN
   Read(Pattern, a);
   CASE a OF
     '#': Read(Pattern, Last);
     '!': Read(Pattern, Threshold);
     '=': JustPvalues:= true;    {invokes Calculate procedure}
     ':': Read(Pattern, Dx);
     '-': Read(Pattern, Length);
     ' ': ;
     '?': BEGIN
  {Element records, Elm[i], as a set of residues, s1 with a molar
   frequency of x, which are to be sought in the range n1 to c1
   from the first element. The frequency of the element, f, is
   calculated as a binomial probability. For a composite element,
   the second component is defined in a similar way by the terms
   s2, n2, and c2; f is then calculated as a joint binomial
   probability.}
        j:= 0; k:= 0;
        x:= 0.0; y:= 0.0;
        simple:= true;
        Read(Pattern, i);
        Read(Pattern, Elm[i].n1);
        REPEAT  Read(Pattern,a);  UNTIL (a = '[');
        Read(Pattern, a);
        WHILE (a <> ']')
        DO BEGIN
          IF (a IN Capitals)
          THEN BEGIN
            Elm[i].s1:= Elm[i].s1 + [a];
            x:= x + Freq[a];
           END;
        Read(Pattern, a)
         END; {WHILE}
        Read(Pattern, Elm[i].c1);
```

```
        WHILE not eoln(Pattern)
        DO BEGIN
          Read( Pattern, a);
          IF (a = '<')              {* if element is composite *}
          THEN BEGIN
            simple:= false;
            Read(Pattern, Elm[i].n2);
            REPEAT  Read(Pattern, a)  UNTIL (a = '[');
            Read(Pattern, a);
            While (a <> ']')
            DO BEGIN
              Elm[i].s2:= Elm[i].s2 + [a];
              y:= y + Freq[a];
              Read(Pattern, a)
             END; {WHILE}
            Read(Pattern, Elm[i].c2),
           END; {IF}
         END; {WHILE}
        IF simple
        THEN BEGIN
          Elm[i].s2 := []; Elm[i].n2:= 0; Elm[i].c2:= 0;
         END
        ELSE BEGIN
               IF (Elm[i].n2 * Elm[i].c2 > 0)
               THEN j:= Elm[i].c2 - Elm[i].n2 +1
               ELSE j:= Elm[i].c2 - Elm[i].n2;
               x := x * (1- Pwr(1-y, j))
             END; {IF}
        k:= Elm[i].c1 - Elm[i].n1 + 1;
        Elm[i].f := 1 - Pwr( 1 - x, k);
        END; {CASE OF '?'}
    END; {CASE}
    Readln(Pattern);
  END; {WHILE}
END; {GetElements}
```

```
FUNCTION IntegerValueOf( Sort: MatchString): integer;
VAR  i, j: SubInt;
     k: integer;
BEGIN
  k:= 1000000;
  FOR i:= 2 TO Last
  DO IF (Sort[i] = '1')
     THEN BEGIN
       j:= i-2;
       k:= k + (2 + (j mod 3)) * round( Pwr(10, j div 3)),
      END; {IF}
  IntegerValueOf:= k;
END; {IntegerValueOf}

PROCEDURE AddToRankList ( Pn: Rankings);
VAR  Cursor, NP: RankPointer;
BEGIN
  new( Cursor);
  Cursor:= Tail;
  REPEAT
    Cursor:= Cursor^.More
  UNTIL (Cursor^.Less^.S <= Pn.S) AND (Cursor^.S >=
Pn.S);
  new(NP);
  NP^.S:= Pn.S;
  NP^.Tag:= Pn.Tag;
  NP^.More:= Cursor;
  NP^.Less:= Cursor^.Less;
  Cursor^.Less^.More:= NP;
  Cursor^.Less:= NP;
  dispose( Cursor);
END;
```

```
FUNCTION Increment (VAR Sort: MatchString): boolean;
VAR  ans, notfound: boolean;
     i, j: SubInt;
BEGIN
  i:= 2;
  ans:= false;
  notfound:= true;
  WHILE NOT ans
  DO IF (Sort[i] = '0')
     THEN BEGIN
       notfound:= false;
       Sort[i]:= '1';
       FOR j:= (i-1) DOWNTO 2
       DO Sort[j]:= '0';
       ans:= true
      END {THEN}
     ELSE IF (i < Last)
          THEN i:= i+1
          ELSE ans:= true;
  Increment:= NOT notfound;
END; {Increment}

PROCEDURE GenerateCombinations;
VAR  Ptrn: Rankings;
     i, j: SubInt;
     UnFinished: boolean;
BEGIN
  Ptrn.S:= 0.0;
  Sort[1]:= '1';
  FOR i:= 2 to Last
  DO Sort[i]:= '0';
  UnFinished:= true;
  WHILE UnFinished
  DO BEGIN
      Ptrn.S:= 1.0;
      FOR i:= 1 TO Last
      DO CASE Sort[i] OF
           '1': Ptrn.S:= Ptrn.S * Elm[i].f;
           '0': Ptrn.S:= Ptrn.S * (1 - Elm[i].f)
         END; {CASE}
      Ptrn.Tag:= IntegerValueOf( Sort);
      IF (Ptrn.S < Threshold)
      THEN AddToRankList( Ptrn);
    UnFinished:= Increment( Sort)
   END; {WHILE}
END; {Generate...}
```

```
PROCEDURE SumRankList (VAR Origin: RankPointer);
VAR Adder: RankPointer;
BEGIN
  new(Adder);
  Adder:= Origin;
  WHILE (Adder^.More^.More <> nil)
  DO BEGIN
    Adder:= Adder^.More;
    Adder^.S:= Adder^.S + Adder^.Less^.S;
   END;
END; {SumRankList}

PROCEDURE AddToTree (VAR Current: RankPointer;
                         NewOne: RankPointer);
BEGIN
  IF (Current = nil)
  THEN BEGIN
    new(Current);
    Current^.Tag:= NewOne^.Tag;
    Current^.S := NewOne^.S;
    Current^.Less := nil;
    Current^.More := nil
   END {THEN}
  ELSE IF (NewOne^.Tag < Current^.Tag)
       THEN AddToTree( Current^.Less, NewOne)
       ELSE IF (NewOne^.Tag > Current^.Tag)
            THEN AddToTree( Current^.More, NewOne);
END; {AddToTree}

PROCEDURE BuildTagTree;
VAR  Cursor: RankPointer;
BEGIN
  Root:= nil;
  Cursor:= Tail;
  REPEAT
    Cursor:= Cursor^.More;
    Tail^.More := nil;
    AddToTree( Root, Cursor);
    dispose( Cursor^.Less)
  UNTIL ( Cursor^.More^.S > Threshold) OR
                            (Cursor^.More^.More = nil);
  REPEAT
    Cursor:= Cursor^.More;
    dispose( Cursor^.Less)
  UNTIL ( Cursor^.More = nil);
END; {BuildTagTree}
```

```
FUNCTION Present (VAR Datum: DataPad; i: SubInt): boolean;
VAR  Result: boolean;
     j, k: SubInt;
BEGIN
  Result:= false;
  IF (Elm[i].s2 = [])
  THEN
   BEGIN
    FOR j:= (Datum.Pos + Elm[i].n1) TO (Datum.Pos + Elm[i].c1)
    DO IF (Seq[j] IN Elm[i].s1)
       THEN  Result:= true;
   END
  ELSE
    FOR j:= (Datum.Pos + Elm[i].n1) TO (Datum.Pos + Elm[i] c1)
    DO IF (Seq[j] IN Elm[i].s1)
       THEN FOR k:= (j + Elm[i].n2) TO (j + Elm[i].c2)
            DO IF (Seq[k] IN Elm[i].s2)
               THEN   Result:= true;
  Present:= Result;
END; {Present}

FUNCTION Significant(VAR Datum: DataPad;
                         Current: RankPointer): boolean;
VAR  ItsFound: boolean;
BEGIN
  ItsFound:= false;
  REPEAT
    IF (Datum.Tag < Current^.Tag)
    THEN Current:= Current^.Less
    ELSE IF (Datum.Tag > Current^.Tag)
         THEN Current:= Current^.More
         ELSE BEGIN
           ItsFound:= true;
           Datum.S := Current^.S
          END;
  UNTIL (Current = nil) OR ItsFound;
 Significant:= ItsFound;
END; {Significant}
```

```
PROCEDURE AddToOrderedList ( Datum: DataPad);
VAR  Cursor, ND: DataPointer;
BEGIN
  new(Cursor);
  Cursor:= Best;
  REPEAT
    Cursor:= Cursor^.More
  UNTIL (Cursor^.Less^.S <= Datum.S)
                AND (Cursor^.S >= Datum.S);
  new(ND);
  ND^.Name:= Datum.Name;
  ND^.Tag:= Datum.Tag;
  ND^.Total:= Datum.Total;
  ND^.Pos:= Datum.Pos;
  ND^.S:= Datum.S;
  ND^.More := Cursor;
  ND^.Less := Cursor^.Less;
  Cursor^.Less^.More := ND;
  Cursor^.Less := ND;
  dispose(Cursor);
END; {AddToOrderedList}

PROCEDURE PrintSequence ( Datum: DataPad);
VAR  i: SubInt;
     TotalP: real;
BEGIN
  FOR i:= 1 to 50
  DO Write( DayOut, Datum.Name[i]);
  Write( DayOut, '   #:', Datum.Total:5);
  Write( DayOut, '; @:', Datum.Pos:4, ', ', Datum.Tag:6),
  Write( DayOut, ',  S: ', Datum.S:9);
  TotalP:= 1 - Pwr((1 - Datum.S),(Datum.Total - Length));
  Write( DayOut, ', P: ', TotalP:9);
  Writeln( DayOut);
  FOR i:= 1 to 130
  DO IF ((Dx-i) mod 10 <> 0)
     THEN Write( Dayout, '.')
     ELSE Write( Dayout, ((abs(Dx-i) div 10) mod 10):1);
  Writeln( DayOut);
  FOR i:= (Datum.Pos + (1-Dx)) TO (Datum.Pos + (130-Dx))
  DO Write( DayOut, Seq[i]);
  Writeln( DayOut);
  FOR i:= (Datum.Pos + (131-Dx)) TO (Datum.Pos + (260-Dx))
  DO Write( DayOut, Seq[i]);
  Writeln( DayOut);
  Writeln( DayOut);
END; {PrintSequence}
```

```
PROCEDURE ScanSequence (VAR Datum: DataPad);
VAR  i, k: SubInt;
     Datum0: DataPad;
     Sort: MatchString;
BEGIN
  Datum0:= Datum;
  FOR i:= 1 TO Datum.Total
  DO BEGIN
    IF (Seq[i] in Elm[1].s1)
    THEN BEGIN
      Sort[1]:= '1';
      Datum.Pos := i;
      FOR k:= 2 TO Last
      DO  IF Present( Datum, k)
          THEN Sort[k] := '1'
          ELSE Sort[k] := '0';
      Datum.Tag:= IntegerValueOf( Sort);
      IF Significant( Datum, Root)
      THEN BEGIN
        AddToOrderedList( Datum);
        PrintSequence( Datum);
       END; {THEN}
      Datum:= Datum0;
     END; {IF}
   END; {FOR}
END; {ScanSequence}
```

```
PROCEDURE NextSequence;
VAR  Size, Kind, NTTL, NTXT, NSEQ, i, j, k: SubInt;
     Res: char;
     Datum: DataPad;
BEGIN
  Datum:= Datum00;
  FOR i:= -260 TO 6000  DO Seq[i]:= ' ';
  FOR i:= 1 TO 10
  DO Read( DayIn, Res);
  Read( DayIn, Size);
  FOR i:= 1 TO 15
  DO Read( DayIn, Res);
  Read( DayIn, Kind, NTTL, NTXT, NSEQ);
  Readln( DayIn);
  IF (Size < Length)
  THEN FOR i:= 1 TO (NTTL+NTXT+NSEQ)   DO Readln( DayIn)
  ELSE BEGIN
    Datum.Total := Size;
    FOR i:= 3 TO 50
    DO BEGIN
      Read( DayIn, Res);
      Datum.Name[i] := res
     END;
    FOR i:= 1 TO (NTTL+NTXT)  DO Readln( DayIn);
    i:= 1;
    FOR j:= 1 TO NSEQ
    DO BEGIN
      FOR k:= 1 TO 36
      DO BEGIN
        Read( DayIn, Res);
        Read( DayIn, Res);
        Seq[i]:= Res;
        i:= i+1
       END; {k-Loop}
      Readln( DayIn);
     END; {j-Loop}
    ScanSequence( Datum)
   END; {ELSE}
END; {NextSequence}

PROCEDURE SearchDayhoff;
VAR i, j, k: SubInt;
    Res: char;
BEGIN
  Reset( DayIn);  ReWrite( DayOut);
  WHILE NOT eof( DayIn)
  DO BEGIN
    Read( DayIn, Res);
    IF (Res = '>')
    THEN NextSequence
    ELSE Read( DayIn, Res);
   END; {WHILE}
END; {SearchDayhoff}
```

```
PROCEDURE PrintRankedListing;
VAR  a: char;
     i: SubInt;
     TotalP: real;
     Cursor: DataPointer;
BEGIN
  Rewrite( RankFile);
Writeln(RankFile,'   Search of NBRF Protein Sequence Database');
  FOR i:= 1 to 3 DO  Writeln( RankFile);
Writeln(RankFile,'No.     Frequency     Element Description');
Writeln(RankFile,'---     ---------     -------------------'),
  FOR i:= 1 TO Last
  DO BEGIN
      Write(Rankfile,i:2,'     ',Elm[i].f:10, Elm[i].n1:8,' ['),
      FOR a:= 'A' to 'Z'
      DO IF (a in Elm[i].s1)
         THEN Write( Rankfile, a);
      Write ( RankFile, '] ', Elm[i].c1:1);
      IF (Elm[i].s2 <> [])
      THEN BEGIN
        Write( RankFile, '  <  ', Elm[i].n2:1, ' [');
        FOR a:= 'A' to 'Z'
        DO IF (a in Elm[i].s2)
           THEN Write( Rankfile, a);
        Write ( RankFile, '] ', Elm[i].c2:1);
       END;
     Writeln( Rankfile);
   END;
  Writeln( RankFile);
  Writeln( RankFile, '   S-value cut-off: ', Threshold:9);
  FOR i:= 1 TO 5
  DO Writeln( RankFile);
  new(Cursor);
  Cursor:= Best;
  WHILE (Cursor^.More^.More <> nil)
  DO BEGIN
    Cursor:= Cursor^.More;
    FOR i:= 1 to 35
    DO Write( RankFile, Cursor^.Name[i]);
    Write( RankFile, ' #:', Cursor^.Total:5);
    Write( RankFile, '; @:', Cursor^.Pos:5);
    Write( RankFile, '; S:', Cursor^.S:9);
    TotalP:= 1 - Pwr((1 - Cursor^.S),(Cursor^.Total - Length)),
    Write( RankFile, '; P:', TotalP:9);
    Writeln( RankFile);
   END; {WHILE}
END; {PrintRankedListing}
```

```
FUNCTION Pvalue ( Current: RankPointer; Tag: integer).real;
VAR  ItsFound: boolean;
     S: real;
BEGIN
  S:= -1.0;
  ItsFound:= false;
  REPEAT
    IF (Tag < Current^.Tag)
    THEN Current:= Current^.Less
    ELSE IF (Tag > Current^.Tag)
         THEN Current:= Current^.More
         ELSE BEGIN
           ItsFound:= true;
           S := Current^.S
          END
  UNTIL (Current = nil) OR ItsFound;
 Pvalue:= S;
END; {Pvalue}

PROCEDURE Calculate;
VAR  a: char;
     i, Tag: integer;
     Flag: PACKED ARRAY [1..20] of char;

BEGIN
     Flag:= '12345678901234567890';
     PrintRankedListing;
     Write( 'Enter 1 or 0 for the presence');
     Writeln( ' or absence of an element.');
     Writeln( 'Quit by entering 0 as first number.');
     WHILE (Sort[1] = '1')
     DO BEGIN
       FOR i:= 1 TO Last
       DO Write( Flag[i]);
       Writeln;
       FOR i:= 1 TO Last
       DO BEGIN
         Read( a);
         Sort[i]:= a
        END;
       Readln;
       Tag:= IntegerValueOf( Sort);
       Writeln( '                    ',Pvalue(Root, Tag):10);
      END; {WHILE}
   END; {Calculate}
```

```
BEGIN

  Dx:= 60;            { default parameters for database search}
  Length:= 0;
  Threshold:= 0.0001;
  JustPvalues:= false;

{******** Average Molar Composition of Proteins ************
 ************************************************************}
  Freq['A']:=0.0815;  Freq['C']:=0.0217;  Freq['D']:=0.0513;
  Freq['E']:=0.0585;  Freq['F']:=0.0391;  Freq['G']:=0.0745;
  Freq['H']:=0.0237;  Freq['I']:=0.0499;  Freq['K']:=0.0638,
  Freq['L']:=0.0884;  Freq['M']:=0.0227;  Freq['N']:=0.0421;
  Freq['P']:=0.0489;  Freq['Q']:=0.0387;  Freq['R']:=0.0471;
  Freq['S']:=0.0721;  Freq['T']:=0.0609;  Freq['V']:=0.0661;
  Freq['W']:=0.0140;  Freq['Y']:=0.0330;
 {************************************************************}

  new(Head);              new(Tail);
  Head^.S:= 1.0;          Tail^.S:= 0.0;
  Head^.More:= nil;       Tail^.More:= Head;
  Head^.Less:= Tail;      Tail^.Less:= nil;
  {Initializes the ranked, doubly-linked list of possible
   pattern matches.}

  new(Best);              new(Worst);
  Best^.S:= 0.0;          Worst^.S:= 1.0;
  Best^.More:= Worst;     Worst^.More:= nil;
  Best^.Less:= nil;       Worst^.Less:= Best ;
  {Initializes the ranked list of pattern matches observed in
   the search of the database.}

  WITH Datum00  { initializes scratchpad     }
  DO BEGIN      { for collecting search data }
    Total:= 0;
    Pos:= 0;
    Tag:= 0;
    S:= 0.0;
   END; {WITH}
  Datum00.Name[1]:= '>';
  FOR i:= 2 to 50  DO Datum00.Name[i]:= ' ';

  GetElements;           { Generates the cumulative probability }
  GenerateCombinations; { distribution associated with the     }
  SumRankList( Tail);    { input pattern and builds the tree of }
  BuildTagTree;          { pattern matches.                     }

  IF JustPvalues
  THEN  Calculate
  ELSE BEGIN
    SearchDayhoff;
    PrintRankedListing
   END;
END.
```