Finding Relevant Biomolecular Features

Lawrence Hunter
Lister Hill Center
Building 38A, MS-54
National Library of Medicine
Bethesda, MD 20894
hunter@nlm.nih.gov

Teri Klein
Computer Graphics Laboratory
10th Floor, Medical Sciences Building
University of California, San Francisco
San Francisco, CA 94143-0446
klein@cgl.ucsf.edu

Abstract

Many methods for analyzing biological problems are constrained by problem size. The ability to distinguish between relevant and irrelevant features of a problem may allow a problem to be reduced in size sufficiently to make it tractable. The issue of learning in the presence of large numbers of irrelevant features is an important one in machine learning, and recently, several methods have been proposed to address this issue. A combination of machine learning approaches and statistical analysis methods can be used to identify a set of relevant attributes for currently intractable biological problems. We call our framework F/I/E (Focus-Induce-Extract). As an example of this methodology, this paper reports on the identification of the features of mutations in collagen that are likely to be relevant in the bone disease *Osteogenesis imperfecta*.

1. Introduction

Biomolecules are complex in both structure and function. Unraveling the relationships between the structures of these molecules and their functions is a core task of modern biology. The complexity of both structure and function makes elucidating these relationships particularly challenging. Using a computational metaphor, the space of possible structures and possible functions are both large; and the space of possible relationships mapping from one space to the other is proportional to the product of an exponential in the size of each! Fortunately, biological research isn't done by exhaustively searching this vast space of all possible relationships between structure and function.

The problems facing researchers looking for relationships between the structure and function of biomolecules are similar to those raised in recent machine learning research under the rubric of "focus of attention." Many inference problems have the characteristic of trying to find a significant pattern or relationship in the presence of many irrelevant features. This problem was recently recognized as a central one in building human-like machine learning systems (Hunter, 1990), and several algorithms have recently been proposed to improve the performance of existing inferential systems in the presence of irrelevant features (Almuallin & Dietterich, 1991; Almuallin & Dietterich, 1992; Kira & Rendell, 1992).

We explored the application of these new machine learning methods to elucidation of structure function relationships in biomolecules. This paper describes a particular application, understanding the structural features related to the severity of the bone disease *Osteogenesis Imperfecta*, but is also intended to provide a general method for the identification of relevant features of biomolecules.

1.1 Collagen and *Osteogenesis Imperfecta*

Collagen is a ubiquitous molecule that provides the tensile strength in the connective tissue of all multicellular organisms. It is the most abundant protein in mammals, constituting a quarter of their total weight. It is the major fibrous element of skin, bone, tendon, cartilage, blood vessels and teeth. There are nearly a dozen different types of collagen found in human beings, coded for by several dozen genes (Stryer, 1988).

Osteogenesis Imperfecta, also known as OI or brittle bone disease, is an genetically transmitted disease of type I fibrillar collagen. The disease is divided into four classes based on its severity, ranging from lethal in the perinatal period (either before or shortly after birth) to nearly asymptomatic. OI type I typically involves little or no deformity, with normal stature, blue sclerae and some hearing loss. OI type III involves short stature, with progressively deforming bones; OI type IV is similar, but less severe. OI type II, however, is lethal. The cause of the disease is mutation in the COL1A1 or COL1A2 genes. These genes code for the $\alpha 1$ and $\alpha 2$ chains of collagen type I. At this point in time, 70 different point mutations in COL1A1 or COL1A2 are known to cause OI, 35 of which cause the lethal form , and 35 of which cause one of the nonlethal forms (Byers, 1990; Byers, personal communication). OI can also be caused by other genetic abnormalities, primarily deletions in the collagen sequence. These other mechanisms will not be discussed here.

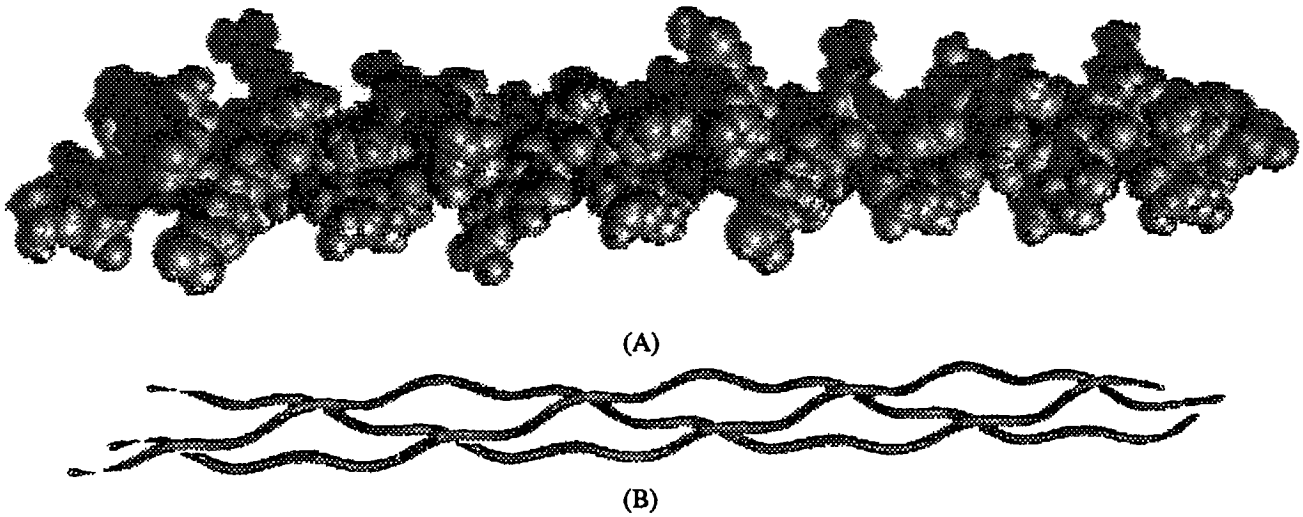

Figure 1: A model of the structure of collagen. This model is derrived by substituting the actual side chains of a 10 residue long region (from 823-832 of the α1 chain) into the crystal structure of a [GLY-PRO-PRO]$_{10}$ polymer and minimizing the energy of the resulting structure. (A) shows a space-filling model of the region; glycine residues form the central core of the chain. (B) shows a ribbon model of the Cα backbone, with glycines depicted in black; this model clearly shows the triple helix. Graphics generated by the UCSF MIDAS program.

Many aspects of the structure of the collagen molecule are known. The molecule is one of the longest known proteins, nearly 3000Å long and about 15Å in diameter. Type I collagen consist of two copies of the α1 chain and one strand of the α2 chain, wound around each other to form a structure that includes a very strong triple helix. Collagen is unusual in several respects. It contains two amino acids that are found in almost no other proteins: hydroxyproline and hydroxylysine. The hydroxylation of the proline and lysine residues occurs as a posttranslational modification of the protein, and only occurs at residues on the amino side of a glycine. Another striking feature of the molecule is its unusual periodic sequence: glycine is nearly every third residue, and the sequence glycine-proline-hydroxyproline occurs repeatedly. Collagen is also unlike most other proteins in that it has an essentially linear conformation, and there are no structural interactions between amino acids that are distant in the sequence. This is very important since it drastically reduces the number of features that must be considered. Figure 1 portrays some of the important features of type I collagen.

Despite the knowledge of the structure of the molecule, and the ability to identify mutations that are both lethal and nonlethal, the precise relationship between the features of a mutation and whether or not it will be lethal remains obscure. For example, there are several mutations to a particular amino acid (e.g. from glycine to serine) that are lethal in one location, and nonlethal in another. This problem of understanding the relationship between a mutation and its phenotype is a common subtask of finding structure/function relationships. The application goal of this paper is to use machine learning methods to discover the factors that influence whether a collagen mutation will be lethal or not.

1.2 Finding Relevant Features

The first step in addressing an induction problem is generally to define the space of features that will be used to describe the examples. This task is challenging because it is constrained by two competing forces. The first constraint is the requirement to include all relevant features. Any information about the objects to be classified not included in the input features cannot be used by any induction method. The competing constraint is the need to keep the number of features used to describe the objects relatively small. There are two reasons for this constraint. First, the computational complexity of induction algorithms generally scales poorly in the number of features used to describe each object. If the algorithm considers all possible combinations of features, then its running time is likely to grow exponentially in the number of features to consider. Second, if there are a large number of features considered by the induction algorithm, the probability that a feature or combination of features will correlate with the distinction being induced simply by chance increases. That is, the possibility of overfitting the data grows with the number of features. This probability

depends both on the number of features and the number of examples, and is especially acute in problems with scarce training data.

In our particular problem, as with many others, it is impossible to depend on existing background knowledge to pick an appropriate sized collection of features to use for induction. There are many possibly relevant features of the collagen molecule that might be involved in whether or not a mutation is lethal. In order to investigate these possibilities, previous work has built simplified models of local regions of collagen and used molecular dynamics to explore the ramifications of particular mutations (Huang, et al., 1991). The model used was based on a X-ray structure of a sequence of ten repeats of the amino acid triple glycine-proline-proline, substituting an actual collagen sequence. A single glycine mutation (to aspartic acid) was explored by energy minimization techniques and molecular dynamics. Although this simulation showed the mutation caused disruptions of the helix, the method is too computationally intensive to be able to build a detailed understanding of the structure/function relationship without more specific direction.

The next step was an attempt at manual discernment of patterns in the collagen sequence around the mutations. Patterns involving the basic biophysical features of charge and hydrophobicity were generated by domain experts and tested on the known mutations. None of these expert-generated patterns were adequate for reliable discrimination of lethal mutations, and the space of possible patterns was recognized to be enormous. For example, note that (Byers, 1990) proposes that the lethality of a mutation is determined by how near it is to the carboxy terminus, and that the extent of this critical region may be determined by the substituting amino acid. In the intervening three years, the number of known mutations has almost doubled, and our results suggest this hypothesis is no longer supported.

The advent of machine learning methods that can handle large numbers of irrelevant features, it has become possible to meaningfully search the entire large space of patterns. The remainder of this paper describes the application of a novel combination of machine learning techniques to the problem.

2. Methods

For each mutation, we know which amino acid was substituted in the mutation, where the mutation occurred in the sequence (including which chain) and whether or not it was lethal (see Table 1, next page). For each of these mutation locations, we can easily deduce a large number of features of its environment.

We propose a Focus-Induce-Extract (F/I/E) method for identifying relevant features. The process involves three stages: first, identify all possibly relevant features, without regard to the problems of efficiency or overfitting in induction. Second apply an focused induction method designed to handle large numbers of irrelevant features. The induction must take place in a framework that allows the measurement of the statistical significance of the results. It is when the results of induction are statistically significant in the aggregate, but do not contain enough information to generate a single set of reliable rules or decision tree that our method is useful. In that case, we decompose the results of the induction to identify the features most likely to be involved in the target relationship.

2.1 The Universe of Possibly Relevant Features

Since the structure of collagen is a linear helix, amino acids far away from the mutation in the sequence will also be far away in the structure. In identifying features of the molecular environment, we need only consider the aspects of the amino acids that are nearby in the collagen sequence, or in the corresponding region on the other chain. The distances at which biochemical forces are effective limits the region of interest to about two turns of the helix in either direction, or about 7 amino acids. To be conservative, we defined the local neighborhood to include three turns of helix (10 amino acids) in either direction from the mutation, on both chains, a total of 42 amino acids. For each of these amino acids, we consider five biochemical features: which amino acid it is, the hydrophobicity, molecular weight, charge and pK (that is, the pH at which the residue becomes charged, if it becomes charged at all; this is set to -1 if it does not become charged). Hydrophobicity was taken from (Cornette, et al., 1987), table 4 (the PRIFT scale) and the other biophysical constants were taken from (Schulz & Schirmer, 1979), table 1-1. In addition, a feature indicating whether the position would be off the end of the chain is noted for each neighbor, and there are three global features for each mutation: which chain it occurred in, the position in the chain (counting from the amino terminus), and whether it is in a proposed critical region, the 60 residues at the carboxy terminus of the chain.[1] So, for each mutation, there are 6*42+3 or 258 features. Most of these features

[1] This critical region is defined by a statistically significant increase in the number of lethal mutations found. Although many nonlethal mutations are also found in this region, the number of lethal mutations found there is greater than the number than would be expected if the mutations were distributed at random, at the $p<.05$ level. Counting from the amino terminus, this region begins at residue 950 and continues to the carboxy end of the chain. There is no corresponding region where there is a greater that expected number of nonlethal mutations.

OI Type II (Lethal)

Ala: 928α1
Ser: 565α1, 598α1, 631α1, 913α1, 964α1, 1003α1, 1009α1, 865α2
Cys: 244α1, 472α1, 691α1, 718α1, 748α1, 904α1, 988α1, 472α2, 787α2
Arg: 298α1, 391α1, 550α1, 667α1, 847α1, 976α1, 496α2, 694α2
Val: 256α1, 637α1, 973α1, 1006α1
Asp: 97α1, 541α1, 559α1, 673α1, 883α1

OI Type I, III or IV (Non-lethal)

Ser: 352α1, 382α1, 541α1, 589α1, 832α1, 844α1, 862α1, 883α1, 901α1, 661α2
Cys: 19α1, 43α1, 46α1, 94α1, 175α1, 178α1, 205α1, 223α1, 382α1, 415α1, 526α1, 864α1, 987α1, 259α2, 646α2
Arg: 79α1, 85α1, 154α1, 1012α2
Val: 586α2
Asp: 547α2, 580α2, 805α2, 907α2, 976α2

GPMGPSGPRGLPGPZGAZGPQGPQGPZGEZGEZGASGPMGPRGPZGPZGKNGDDGEAGKZGRZGERGPZGPQGARGLZGTAGLZGMKGHRGFSGLDGAKGDAGPAGPKGEZGSZGENGAZGQMG
PRGLZGERGRZGAZGPAGARGNDGATGAAGPZGPTGPAGPZGPZGAVGAKGEAGPQGPRGSEGPQGVRGEZGPZGPAGAAGPAGNZGADGQZGAKGANGAZGIAGAZGPZGARGPSGPQGZGGP
ZGPKGNSGEZGAZGSKGDTGAKGEZGPVGVQGPPGPAGEEGKRGARGEZGPTGLZGPZGERGGPGSRGFZGADGVAGPKGPAGERGSZGPAGPKGSZGEAGRZGEAGLZGAKGLTGSZGSZGPDG
KTGPZGPAGQDGRZGPZGPZGARGQAGVMGFZGPKGAAGEZGKAGERGVZGPZGAVGPAGKDGEAGAQGPZGPAGPAGERGEQGPAGSPGPQGLZGPAGPZGEAGKZGEQGVZGDLGAPGPSGA
RGERGFZGERGVQGPZGPAGPRGANGAZGNDGAKGDAGAZGAZGSQGAZGLQGMPGERGAAGLZGPKGDRGDAGPKGADGSZGKDGVRGLTGPIGPZGPAGAZGDKGESGPSGPAGPTGARGA
ZGDRGEZGPZGPAGFAGPZGADGQZGAKGEZGDAGAKGDAGPZGPAGPAGPPGPIGNVGAZGAKGARGSAGPZGATGFZGAAGRVGPZGPSGNAGPZGPZGPAGKEGGKGPRGETGPAGRZGEV
GPZGPZGPAGEKGSZGADGPAGAZGTZGPQGIAGQRGVVGLZGQRGERGFZGLZGPSGEZGKQGPSGASGERGPZGPMGPZGLAGPZGESGREGAZGAEGSZGRDGSPGAKGDRGETGPAGPZGAZ
GAZGAZGPVGPAGKSGDRGETGPAGPAGPVGPAGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPZGSZGEQGPSGASGPAGPRGPZGSAGAPGKDGLNGLZGPIGPZGPRGRTGDA
GPVGPPGPZGPZGPZGPP

GPMGLMGPRGPZGAAGAZGPQGPQGPAGEPGEZGQTGPAGARGPAGPZGKAGEDGHZGKPGRZGERGVVGPQGARGPFZGTZGLZGFKGIRGZNGLDGLKGQZGAZGVKGEZGAZGENGTPGQTG
ARGLZGERGRVGAZGPAGARGSDGSVGPVGPAGPNGSAGPZGFZGAZGPKGEIGAVGNAGFTGPAGPRGEVGLZGLSGPVGPZGNPGANGLTGAKGAAGLZGVAGAZGLZGPRGIPGPZGAAGTT
GARGLVGEZGPAGSKGESGNKGEZGSAGPQGPZGPSGEEGKRGPNGEAGSAGPZGPZGLRGSZGSRGLZGADGRAGVMGPPGSRGASGPAGVRGPNGDAGRZHEZGLMGPRGLZGSZGNIGPAGK
EGPVGLZGIDGRZGPIGPVGARGEZGNIGFZGPKGPTGDZGKNGDKGZAGLAGARGAZGPDGNNGAQGPZGPQGVQGGKGEQGPAGPZGPQGLZGPSGPAGEVGKPGERGLHGEPGLZGPAGPRGE
RGPZGESGAAGPTGPIGSRGPSGPZGPDGNKGEZGVVGAVGTAGPSGPSGLZGERGAAGIZGGKGEKGEPGLRGEIGNZGRDGARGAHGAVGAZGPAGATGDRGEAGAAGPAGPAGPRGSZGERGE
VGPAGPNGFAGPAGAAGQZGAKGERGGKGPKGENGVVGPTGPVGAAGPAGPNGPZGPAGSRGDGQPZGMTGFZGAAGRTGPZGPSGISGPZGPPGPAGKEGLRGPRGDQGPAGRZGEVGAZGPPG
FAGEKGPSGEAGTAGPZGTZGPQGLLGAPGILGLZGSRGERGLZGVAGAVGEZGPLGIAGPZGARGPZGAVGSZGVNGAZGEAGRDGNPGNDGPZGRDGQZGHKGERGYHGNIGPVGAAGAZGPZ
GPVGPAGKZGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEZGEKGPRGLPGFKGHNGLQGLZGIAGHHGDQGAZGSVGPAGPRGPAGPSGPAGKDGRTGHZGTVGPAGIRGPQCHQGPAGPZ
GPZGPZGPZGVS

Table 1: The mutations in collagen that lead to OI. Each mutation is represented by a number representing a position in the chain (counting from the amino terminus) followed by either α1 or α2. The long strings at the bottom of the figure are the sequences of the α1 and α2 chains respectively, using standard amino acid abbreviations, except for Z, which here stands for hydroxyproline.

are numeric (position, hydrophobicity, molecular weight, pK), and the remainder are nominal.

Recently, an explicit method for identifying potentially relevant features given a labeled set of examples was proposed as a way of improving the performance of induction algorithms in the presence of many irrelevant features (Almuallin & Dietterich, 1991; Almuallin & Dietterich, 1992). However, this method, called FOCUS is both computationally inefficient, and performs poorly in the presence of noise. (Kira & Rendell, 1992) propose a statistically based method for identifying probably relevant features called RELIEF. This method has several advantages over FOCUS: It has linear time complexity in the number of features it is considering (compared to FOCUS's, which can have time complexity $O(2^p)$ for p features, if all p turn out to be relevant). RELIEF is also effective in the presence of noise (errors) in the data.[2]

RELIEF generates relevance weights for each feature, and requires as a parameter τ that determines the cutoff above which a feature will be considered relevant. Although Kira and Rendell provide a method for estimating this parameter, the method either requires knowledge of the distribution of the relevance levels for each feature, or generates very large values for τ. They also note that its usually possible to determine an appropriate value for τ by inspection. Some of the features generated by the above exhaustive method are known to be irrelevant. For example, the amino acids ±3, ±6 and ±9 from the mutation in the same chain should almost always be glycine, and the features derived from those positions should be irrelevant. The highest scoring of these features

[2] An implementation of RELIEF in Common Lisp is available from the first author.

had a relevance score of 0.071. Another way of estimating the cutoff was to look for the lowest scoring feature we though a priori was likely to be important. We looked for the lowest relevance score of an attribute of the mutation itself, which turned out to be its pK value, with a relevance score of 0.204. Since there were 83 features at or above the 0.2 level, and 127 features at or above the 0.07, we decided to use the stricter figure as the cutoff.

Once the set of potentially relevant features is selected, they are used to generate decision trees using the C4.5 induction algorithm (Quinlan, 1986; Quinlan, 1987; Quinlan, 1991). In order to evaluate the effectiveness of the induction, it is embedded in a k-way cross-validation approach. The data is randomly divided into K approximately equal sized subsets[3] . The induction program is run K times; in each run, one of the subsets is used as the test set, and the other K-1 subsets are combined to form the training set. Each run produces a separate decision tree. Each tree is tested on the appropriate test subset. If the average performance of the collection of K induced decision trees is significantly different from what would be expected at random, than the induction method is reliably finding a relationship between the input features and the predicted outcome.

There are several free parameters that must be set in these runs. The "m" parameter of C4.5 sets a minimum number of objects in each branch; since there were a relatively small number of examples, we incremented this parameter from the default 2 to 3. We used probabilistic thresholds for continuous attributes (the "p" flag). For the cross validation, we selected k=6 in order to give a reasonable number of instances in each test set (11 or 12). The members of a set of decision trees that performs above the random level will not necessarily be identical. The portions of the trees that are similar to each other are presumably the source of the predictive accuracy of the trees. (Redundancy among the input features, which RELIEF does not detect, may belie this assumption.) The goal of the extract portion of the F/I/E method is to find the similarities among the induced trees.

Three methods to extract similarities among the trees were used. First, we looked for rules implied by more than one decision tree. All decision trees can be syntactically transformed into a set of rules. Such a set contains a rule for each path from the root of the tree to a leaf. The antecedent of each rule is the collection of values implied by the decisions along the path, and the consequent is the value at the leaf. Some paths through a decision tree may have little or no data supporting them, and can be pruned away. The C4.5rules program uses this method to extract rules. We used the most stringent pruning method (the "-P3" flag), which requires that each antecedent in a potential rule pass a statistical significance test. After generating rule sets for each of the K decision trees, we identified rules that appeared in more than one tree. We also manually searched for rules that were semantically related variants of each other.

The second extraction method identified features that appeared in multiple decision trees. For each input feature (e.g. the hydrophobicity of the residue +2 of the mutation), we counted the number decision trees in which that feature appeared. For this calculation, if the pruned tree was more accurate on the unseen data than the unpruned tree, a feature was only counted if was in the pruned tree, since presumably features the additional features in the unpruned tree resulted in overfitting. This resulted in a list of features that appeared in multiple trees.

The third method for extracting similarities was to count the number of times any feature of a particular residue was mentioned in a decision tree. For example, if the feature "hydrophobicity of the residue +2 of the mutation" appeared in one tree, and the feature "charge of the residue +2 of the mutation" appeared in another, the residue +2 of the mutation would be credited with appearing in two trees. Although it is possible to count the number of common rules or features in any set of decision trees, this aggregate count is only possible because of the way the features were generated for this experiment, and it may or may not be possible to duplicate this extraction in other systems.

3. Results

Trees induced over the entire feature set (without using RELIEF) were not statistically different than random guessing. However, using only the relevant features in the instance descriptions resulted in a set of decision trees that were statistically better than random guessing at the $p<.05$ level. The average accuracy of the unpruned trees on unseen test cases was 68.7% correct with a standard deviation of 7.7%. The average performance of the pruned trees was 63.2% correct with a standard deviation of 12.7%, which is not statistically different from random guessing. Using the best trees from each run (4 pruned and 2 unpruned), the average accuracy is 69.2% with a standard deviation of 7.0%, which is significantly different from random.

So, in the aggregate, we can be reasonably certain that C4.5 is finding decision trees that are capable of generalizing to unseen data. However, the 6 trees generated in the cross validation runs are quite different from each other, and there is no statistically sound reason to prefer any one of the trees over any other. It is perfectly plausible that the tree that performed the best on the unseen data in these cross-validation runs was better than the tree

[3] The particular random division can have an effect on the outcome. The division we used is available from the authors.

-2 of mutation is PRO → lethal (3 trees)
Right: 21 (67%) Wrong: 10 (32%)
-2 of mutation is ALA → non-lethal (2 trees)
Right: 9 (75%) Wrong: 3 (25%)
+2 of mutation pK > 10.5 → lethal (2 trees, in 3rd tree with threshold at 6)
Right: 6 (100%) Wrong: 0 (0%) with threshold at 10.5
Right: 8 (80%) Wrong: 2 (20%) with threshold at 6
+4 of mutation charge negative → lethal
Right: 10 (76%) Wrong: 3 (23%)
+4 of mutation charge positive and +2 of mutation pK ≤ 6 → non-lethal
Right: 6 (100%) Wrong: 0 (0%)

Figure 2: The first three rules are extracted from more than one tree. The last pair of rules are related, but not identical; they were extracted from one tree each. The correctness figures are taken over all training and test examples.

that had the worst performance strictly by chance. There is too little data to treat any one of these trees as reliable.

The fact that the induced decision trees are so different from each other is an indication that there are still problems with the data that prevent C4.5 from finding the "true" decision tree. We tried increasing the relevance threshold τ for RELIEF to 0.3, but in that case C4.5 did not find decision trees that were better than chance; we must have eliminated some truly relevant features by using that high a threshold.

We then applied our extraction methods to the set of statistically significant trees that we induced. The first step was to extract rules from each of the six induced trees in the cross validation run, and look for rules that appeared in multiple trees. The results of this extraction are shown in Figure 2. We then examined the features that were used in each of the trees. Each induction run produces both a pruned and an unpruned tree. Pruning removes branches of the tree, reducing complexity that may be a result of overfitting. (This is a different kind of pruning than that done in the rule generation step, above; see (Quinlan, 1987).) For each of the decision trees, selected the version that had the best accuracy on the unseen data; four out of six of these were the pruned versions. We made a histogram of the number of times that each feature appeared in a tree (Figure 3a, next page). The most commonlyoccurring features occurred in three of the six trees; these three trees were also the best performing trees of the six (which is at best mildly suggestive).

We then further aggregated this information to look for which amino acids were the subject of the features in the decision trees. So, for example, any mention of the hydrophobicity, charge, pK, etc. of the amino acid -2 of the mutation would get counted for that amino acid. One amino acid, the one +4 of the mutation, was mentioned in 5 out of 6 trees. Features of four other amino acids were mentioned in at least three of the trees (see Figure 3b, next page). Out of the 42 amino acids originally considered, features from only 5 of them appear consistently in the decision trees.

We also tested if the relevant features we found had different distributions in the mutation data than they do in the unmutated wildtype collagen. For every feature in every tree, we found that the distribution of the feature values for the wildtype and the mutation were statistically indistinguishable; the distributions only became skewed when comparing lethal vs. non-lethal mutations (Figure 4, last page).

4. Discussion

There are several biological surprises in these findings. Most important, the position of the mutation in the chain does not appear to be as relevant in determining its lethality as previously believed. Position gets a fairly low RELIEF score, and does not show up in any of the decision trees. Expert opinion had suggested that this was the main determining feature of lethality (Byers, 1990). We did find a statistically significant increase in lethality at the very end of the $\alpha 1$ chain, and there is a biophysical reason to believe that the carboxy terminus is more sensitive to mutations (it is where the triple helix begins to form). However, our F/I/E results suggest that this is not the main determining factor of lethality. Neither absolute position of the mutation nor its proximity to the carboxy terminus appears to be a good way to distinguish lethal from nonlethal mutations. The hypothesis suggested by these results is that the atomic environment of the mutation is the most important feature in determining its lethality.

A second somewhat surprising finding is the relatively low importance of the mutation itself, compared to its surrounding environment. These results clearly suggest that the features of the amino acids ±2, +4 and +4 on the opposite chain of the mutation are as important as the mutation itself in determining lethality, or perhaps more so.

Appeared in three trees:

chain
+2 of mutation pK
-2 of mutation amino acid

Appeared in two trees

mutation hydrophobicity
+4 of mutation charge
+4 of mutation hydrophobicity
+4 of mutation on opposite chain hydrophobicity
+4 of mutation on opposite chain pK

(A)

Features from this amino acid appeared in five trees

+4 of mutation

Features from these amino acids appeared in four trees

+2 of mutation
+4 of mutation on opposite chain

Features from these amino acids appeared in three trees

-2 of mutation
mutation itself

(B)

Figure 3: (A) The features that appeared in more than one tree. (B) The amino acids whose features appeared in three or more decision trees.

This is a possible explanation for the difficulty in analyzing the model used in (Huang, et al., 1991).

A third implication of this work is in suggesting the ordering of the chains in the triple helix. There are two topologically distinct orderings: the two $\alpha 1$ chains may beadjacent to each other, or separated by the $\alpha 2$; that is, the order could be $\alpha 1, \alpha 1, \alpha 2$ or $\alpha 1, \alpha 2, \alpha 1$. The true ordering is not currently known. Molecular models show that the spatial neighbors across the helix are position i, i+1 and i+2 respectively in the three chains. Making the biophysically reasonable assumption that helical neighbors are important factors in lethality, our finding that the amino acids ± 2 from the mutation are relevant (and the amino acids ± 1 are not) suggests that the chain ordering is $\alpha 1, \alpha 2, \alpha 1$. The reason is that in this case, the helix neighbors of an amino acid in one of the $\alpha 1$ chains is the residue ± 2 of that amino acid in the other $\alpha 1$ chain. If the order were $\alpha 1, \alpha 1, \alpha 2$, then the helix neighbor would be the ± 1 amino acids, and presumably it would have played a role in the lethality predictions.

All of these implications must be considered tentative. There is not enough data yet available to induce a reliable decision tree or rule set for predicting the lethality of unseen mutations. However, our findings may be useful in suggesting more realistic models for analysis with molecular dynamics and other powerful (but compute intensive) tools, and in reducing the number of simulations that need to be done. Our hypotheses also suggest empirical work that could be done.

For the computational science community, the significance of this work is primarily in the demonstration that the F/I/E method makes it possible to productively explore a large feature space with a relatively small amount of data. There are a large number of problems, both in elucidating structure function relationships and elsewhere that can be reasonably cast in this framework. Although we believe the analysis of the features used in induced rules or decision trees is not uncommon among machine learning researchers (although rarely published), we feel that F/I/E's combination of RELIEF/C4.5 and this style of post-induction feature analysis is clearly a novel approach. F/I/E is also an example of the usefulness of the "planning to learn" idea described in (Hunter, 1990), demonstrating the benefits that combinations of multiple inference methods can provide.

There are several limitations of this method. Most important is the original selection of input features. Although the F/I/E method allows for the consideration of a large number of features, it does not guarantee that the correct features will be included in the original set. Of course, if relevant features are omitted from the original representation, the induction will likely fail. Also important is keeping the amount of redundancy in the feature set low. RELIEF is not capable of identifying

	uncharged	negative	positive
wildtype	518 (76%)	109 (16%)	49 (8%)
mutation	48 (68%)	13 (19%)	9 (13%)
lethal mutation	23 (65%)	3 (9%)	9 (26%)
nonlethal mutation	25 (71%)	10 (34%)	0 (0%)

Figure 4: The distributions of the value of the feature **+4 of mutation charge.** The first row shows the baseline distribution of the charge values of the residues +4 from the repeating glycines in unmutated collagen (all OI mutations occur in these positions). The second row shows a very similar distribution holds over the residues +4 from actual mutations. However, the last two rows show that when the mutations are subdivided into lethal and nonlethal, the distribution is quite skewed.

redundant features, and their presence interferes with the effectiveness of the relevant feature extraction. Finally, the F/I/E method is working at the very edge of the ability to make well founded inferences. At best, the suggestions the method makes about relevant features should be taken as plausible hypotheses, in need of additional testing.

Acknowledgments

The authors wish thank Dr. Peter Byers, who provided the mutation data; Edison Wong and Elham Mamoodzadeh-Kashi, who worked on related projects as students in the NLM Medical Informatics Elective program; and Dr. Robert Langridge, Director of the Computer Graphics Laboratory, UCSF. Dr. Klein also acknowledges support from National Institutes of Health grant RR 1081(TK).

References

Almuallin, H., & Dietterich, T. (1991). Learning with Many Irrelevant Features. In *Proceedings of Ninth National Conference on Artificial Intelligence, (AAAI-91)* vol. 2, pp. 547-552. AAAI Press.

Almuallin, H., & Dietterich, T. (1992). Efficient Algorithms for Identifying Irrelevant Features. In *Proceedings of Canadian AI Conference*, (reprint).

Byers, P. H. (1990). Brittle bones, fragile molecules: disorders of collagen gene structure and expression. *Trends in Genetics*, 6(9), 293-300.

Cornette, J., et al (1987). Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins. *Journal of Molecular Biology*, 195, 659-685.

Huang, C. C., et al (1991). The Macromolecular Workbench and its application to the study of Collagen. In *Proceedings of Hawaiian International Conference on System Sciences - 24*, vol. 1, pp. 636-643. IEEE Computer Press.

Hunter, L. (1990). Planning to Learn. In *Proceedings of The Twelfth Annual Conference of the Cognitive Science Society*, (pp. 26-34). Boston, MA:

Kira, K., & Rendell, L. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. In *Proceedings of National Conference on AI (AAAI-92)*, (pp. 129-134). San Jose, CA: AAAI Press.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81-106.

Quinlan, J. R. (1987). Simplifying Decision Trees. *International Journal of Man-Machine Studies*, 27, 221-234.

Quinlan, J. R. (1991). C4.5. Computer program, available from the author: quinlan@cs.su.oz.au.

Schulz, G. E., & Schirmer, R. H. (1979). *Principles of Protein Structure*. New York: Springer-Verlag.

Stryer, L. (1988). *Biochemistry*. New York, NY: W.H. Freeman.