Softening constraints in constraint-based protein topology prediction

Simon Parsons *

Advanced Computation Laboratory,
Imperial Cancer Research Fund
P.O. Box 123, Lincoln's Inn Fields
London WC2A 3PX
email: sp@acl.lif.icnet.uk

Department of Electronic Engineering
Queen Mary and Westfield College
Mile End Road
London E1 4NS

Abstract

This paper is concerned with the handling of uncertain data about the applicability of constraints in protein topology prediction. It discusses the use of novel methods of representing and reasoning with uncertain data, and presents the results of some experiments in using these methods to build probabilistic models of constraint application. It thus builds on work by other authors in both constraint satisfaction and probabilistic reasoning.

Introduction.

In recent years there has been a lot of interest in the use of both constraint satisfaction techniques and probabilistic models for solving problems in molecular biology. Constraint satisfaction techniques have been applied to problems such as RNA structure prediction (Altman 1994a), tertiary protein structure prediction (Altman & Jardetzky 1989), genetic map assembly (Clark, Rawlings, & Doursenot 1994), and protein topology prediction (Clark, Shirazi, & Rawlings 1992; Clark *et al.* 1994). Probabilistic models have also been widely applied. Their use has mainly been restricted to the prediction of protein structure (Asai, Hayamizu, & Onizuka 1993; Brown *et al.* 1994; Delcher *et al.* 1993; 1994; Haussler *et al.* 1993), but they have also been applied to the prediction of RNA structure (Altman 1994b), and tasks such as pedigree analysis (Spiegelhalter 1990; Szolovits 1992). There have even been a few applications which combine constraint satisfaction and the use of probability and similar measures (Altman 1994a; Altman & Jardetzky 1989; Clark *et al.* 1994).

*This work was partially supported by ESPRIT Basic Research Action 3085 (DRUMS).

The reason for this interest is that each technique provides a means of handling one of the more difficult aspects of molecular biology problems—that they involve hugh search spaces, and that they are full of uncertain and noisy information (Allison 1993). The fact that the search spaces are so large means that it is not feasible to search them exhaustively and so some means of restricting the search is required. The "constrain and generate" methods adopted by systems of constraint satisfaction do precisely this. The fact that data is uncertain and noisy suggests that methods that take this into account by explicitly modelling the imperfections in the data (MacKenzie, Platt, & Dix 1993) will be profitable, and this is borne out by the good performance of probabilistic methods for protein structure prediction (Delcher *et al.* 1993; 1994).

In this paper we describe some experiments in explictly representing the uncertainty of a set of constraints used for protein toplogy prediction which could be combined with the constraint-based approach of Clark and colleagues (Clark, Shirazi, & Rawlings 1992) to soften the constraints, making it possible to model situations in which they do not definitely hold. This work uses an alternative model of uncertainty to that described in (Clark *et al.* 1994) which is based upon new means of representing and reasoning with uncertainty known as valuation systems (Shenoy 1991) and the theory of qualitative change (Parsons 1995b). As with previous work in this area, the intention is to help a molecular biologist explore the space of possible protein structures, rather than to present her with a structure that is guaranteed to be correct.

Protein topology prediction.

Since protein function is closely related to structure, protein structure prediction—taking the amino acid

sequence of the primary structure, determining the location of α-helices and β-strands that make up the secondary structure, and then discovering the 3-D position of every atom in the tertiary structure—is an important problem in molecular biology, and one that can be eased by the use of computational methods. Protein topology is an intermediate level somewhere between secondary and tertiary structure which specifies how secondary structural units combine together into larger complexes such as α/β-sheets. The prediction of protein topology is particularly interesting because it can be used to guide the choice of experiments to confirm protein structure and to search for similar known structures, whilst being easier to establish than the full 3-D structure. However, a vast number of possible topologies can be hypothesised for a given secondary structure, for example, a mixed α/β-sheet of n strands, where $n > 1$, can be arranged in $\frac{n!(4n-1)}{2}$ possible ways (Clark *et al.* 1994). One way to reduce this space is to identify and apply constraints based upon previous analyses of similar proteins. For example, for α/β sheets (Clark *et al.* 1994; Taylor & Green 1989) we might use[†]:

- C1. For parallel pairs of β-strands, β-α-β and β-coil-β connections are right handed.
- C2. The first β-strand in the sheet is not at the edge of the sheet.
- C3. Only one change in winding direction occurs.
- C4. The β-strands associated with the conserved patterns lie adjacent in the sheet.
- C5. All strands lie parallel in the β-sheet.
- C6. Unconserved strands are at the edge of the sheet.
- F1. Strands are ordered in the sheet by hydrophobicity, with the most hydrophobic strands central.
- F2. Parallel β-coil-β connections contain at least 10 amino acids.
- F3. Large insertions and deletions are expected to occur on the edge of a domain.
- F4. Most conserved loops lie adjacent in front.
- F5. Long secondary-structure units should lie parallel or antiparallel to one another, with sequential units being antiparallel.

These constraints can be applied manually, as described by Taylor and Green (Taylor & Green 1989), or by generating all possible topologies and removing those that do not conform to the constraints (Cohen & Kuntz 1987). However manual search is a time-consuming and error-prone procedure suitable only for small sheets, and exhaustive search is far too inefficient to be applied to large structures. As a result, Clark and colleagues (Clark, Shirazi, & Rawlings 1992) developed a Prolog program named CBS1 (later re-implemented in ElipSys as CBS1e and CBS2e (Clark *et al.* 1994)) to apply the constraints. In this constraint-based approach, the search proceeds by incrementally adding components (such as β-strands) to a set of possible structures. After each addition the set of structures is pruned by testing against every constraint. CBS1 was used to reproduce Taylor and Green's results as well as to identify a new topological hypothesis consistent with the constraints (Clark, Shirazi, & Rawlings 1992), indicating that the original search was not exhaustive.

As one might imagine, because the constraints are derived from aggregate properties of a collection of proteins, they do not apply to all of them. Clark and colleagues (Clark, Shirazi, & Rawlings 1992) assessed the validity of $C1$, $C2$, $C3$, $C5$, $F1$ and $F2$ by checking them against the known structures of eight nucleotide binding domains with similar function. The results are reproduced in Table 1, where the structures that are grouped together are those relating to the same protein. For instance *p1gpd*, *p1gd1* and *p2gpd* are different experimentally determined structures for D-glyceraldehyde-3-phosphate dehydrogenase. Each of the variations should be considered equally valid, so when a rule holds for one form of a protein and not for another, it is ambiguous whether or not the constraint holds for that protein. Other results support the idea of constraints being uncertain. For instance, King and colleagues (King *et al.* 1994) found that some of the constraints used by Clark et al. failed to hold for some proteins in wider domains.

Thus while the folding rules are useful heuristics, they are only true some of the time, leading us to suspect that explicitly modelling the uncertainty in the constraints might be advisable. One approach to doing this is to assess the validity of a structure based upon the constraints to which that structure conforms, and is one of the subjects of this paper. This was also proposed in (Clark *et al.* 1994) though here we use a more sophisticated model. This paper also explores an alternative method based upon results for propagating

[†]The names of the constraints come from (Taylor & Green 1989) where $C1$–$C6$ are the constraints applied during the construction of the solutions ($C6$ being implicit as argued in (Clark, Shirazi, & Rawlings 1992)), and $F1$–$F5$ are generally applicable "folding rules" used to assess whether predicted structures are valid

Protein ID.	Constraints Violated	Protein ID.	Constraints Violated	Protein ID.	Constraints Violated
p4adh	F1	p3dfr	C3 C5 F1	p1pfk	C5 F1
p5adh	F1	p4dfr	C3 C5 F1 F2	p2pfk	C5 F1
p6adh	F1			p3pfk	C5
p7adh	F1	p3adk		p4pfk	C5
p1ldx		p3grs	C2 F1	p1gpd	C3 C5 F1 F2
p3ldh	F1			p1gd1	C3 C5 F1 F2
p4ldh	F1	p3pgk	F1	p2gpd	C3 C5 F1 F2

Table 1: The results of checking constraints against eight nucleotide-binding domain proteins

Constraint (x)	Number of cases in which the constraint is violated	$p(xA)$
C1	0	1.0
C2	1	0.947
C3	5	0.737
C5	9	0.526
F1	15	0.211
F2	4	0.789

(a)

Constraint (x)	Number of cases in which the constraint is violated	$p(xA)$
C1	0	1.0
C2	1	0.875
C3	2	0.75
C5	3	0.625
F1	5–7	[0.125, 0.375]
F2	1–2	[0.750, 0.875]

(b)

Table 2: Probabilities of constraints holding based upon the (a) "disambiguated" and (b) "pure" interpretations

qualitative changes in probability (Parsons 1995b).

Softening the constraints.

The best way of modelling the uncertainty in the constraints is not clear, and so in the tradition of experimental investigations of the best way of modelling uncertainty in a given problem (Heckerman 1990; Heckerman & Shwe 1993; Saffiotti, Parsons, & Umkehrer 1994) we discuss a number of different ways in which the data from Table 1 may be represented. There are, of course, other possibilities which are not discussed here—we just cover the most obvious probabilistic models—and for methods using other uncertainty handling techniques see (Parsons 1995a). Since the data is drawn from a reasonably random population of proteins the following simple argument can be made. Table 1 holds a list of 8 proteins. Of these, 7 conform to constraint $C2$, and 1 does not, so a possible nucleotide binding domain structure that conforms to $C2$, has a probability of:

$$P(C2A) = \frac{\text{Number of proteins for which } C2 \text{ holds}}{\text{Total number of proteins}} = \frac{7}{8}$$

of being a real protein. Since the sample size is very small, the probabilities will not be very accurate, but they will be the best values that can be obtained given the data to hand.

However, there is a problem with this approach that arises because the data is ambiguous. Of the eight proteins analysed, several have alternative structures and some constraints hold for some alternative structures and not for others. Thus it is not clear whether or not some constraints are valid for some proteins. To handle the ambiguity we need more subtle approaches, and one is to "disambiguate" and consider each of the 19 possible structures as a separate entity. Doing this allows us to argue that of these 19 structures 18 conform to $C2$ and 1 doesn't so that a structure that conforms to $C2$ has a probability of:

$$P(C2A) = \frac{\text{Number of structures for which } C2 \text{ holds}}{\text{Total number of structures}} = \frac{18}{19}$$

of being a real protein. This approach gives the probabilities of Table 2(a).

However, it could be argued that disambiguation distorts the data, and the uncertainty should be modelled in a "purer" way acknowledging the ambiguity. One way of doing this is to use interval probabilities to represent the uncertainty with the lower bound calculated by counting proteins for which the rule is ambiguous as proteins for which it fails to hold, and the upper bound by counting proteins for which the rule is ambiguous as proteins for which it does hold. So, for a constraint for which there is no ambiguity, for instance $C2$, we have, as before:

$$P(C2A) = \frac{\text{Number of proteins for which } C2 \text{ holds}}{\text{Total number of proteins}} = \frac{7}{8}$$

For ambiguous constraints such as $F1$, we use the following. We have one protein for which the constraint is known to hold for all structures and three for which it is known to hold for at least one structure, so:

$$P(C2A) = \left[\frac{\text{Number of proteins for which } C2 \text{ holds for every structure}}{\text{Total number of proteins}}, \frac{\text{Number of proteins for which } C2 \text{ holds for at least one structure}}{\text{Total number of proteins}}\right] = \left[\frac{1}{8}, \frac{3}{8}\right]$$

Using this method on data in Table 1 we get the probabilities of Table 2(b). Other methods of handling the ambiguity may be adopted (Parsons 1995a).

The available data can also be interpreted as telling us how often constraints hold for real proteins, since every structure in the table occurs in nature. Thus the proportion of the proteins for which a given constraint holds is the conditional probability that the constraint holds given that the protein is real. Thus, for $C2$:

$$p(C2|real) = \frac{\text{Number of proteins for which } C2 \text{ holds}}{\text{Total number of proteins}}$$

We have no information about the proportion of proteins for which $C2$ holds yet which are not real, so we cannot establish $p(C2 \mid \neg real)$ in the same way. Instead, we must employ the principle of maximum entropy to conclude that $p(C2 \mid \neg real) = 0.5$. From (Parsons 1995b) we learn that these values are sufficient to establish the relationship between $p(C2)$ and $p(real)$ in terms of the derivative $\frac{dp(C2)}{dp(real)}$ which relates them. This information, in turn (Parsons 1995b), is sufficient to give us $\frac{dp(real)}{dp(C2)}$, allowing us to establish how $p(real)$ changes when we have information about $C2$ holding. From the data we have, irrespective of whether we use the "disambiguated" or "pure" interpretations, we have the derivatives of Table 3. Note that $\frac{dp(real)}{dp(C1)} = [+]$ indicates that as $p(C1)$ increases, so does $p(real)$, and $\frac{dp(real)}{dp(F1)} = [-]$ indicates that as $p(F1)$ increases, $p(real)$ decreases.

This information about changes in probability fits closely to CBS1, since in that system, following each step, a structure can either conform to the same set of constraints as before, or to some superset or subset of that set. So, after each step new evidence about whether or not a constraint holds may be available. If it is possible to relate the fact that a particular structure conforms to a particular constraint to it being correct, then the effect of the new knowledge may be propagated to find out how it affects the likelihood that the structure is correct. Thus it is possible to tell whether the protein structure that is being assembled has become more or less likely to be correct, and whether it should be rejected or continued with accordingly.

Constraint (x)	Number of cases in which the constraint is violated	$\frac{dp(real)}{dp(x)}$
$C1$	0	[+]
$C2$	1	[+]
$C3$	2	[+]
$C5$	3	[+]
$F1$	5–7	[−]
$F2$	1–2	[+]

Table 3: The probabilistic qualitative derivatives based upon both "disambiguated" and "pure" interpretations

The valuation system models.

Having considered the different ways in which the uncertain nature of the constraints can be modelled, we turn to considering how to employ these models in topology prediction. CBS1 generates as its output sets of possible topologies of nucleotide binding domain proteins and the constraints to which they conform. One kind of output that would be useful is some measure of the validity of the sets of topologies based upon the constraints to which they conform.

To build suitable representations, valuation systems were used. Valuation systems (Shenoy 1991) can be viewed as a generalisation of probabilistic networks since they allow a variety of uncertainty handling methods to be employed including evidence theory (Shafer 1976) and possibility theory (Zadeh 1978) whilst losing none of the expressiveness[‡] or computational efficiency of probabilistic networks (Delcher *et al.* 1994). They were selected for this work because they make it possible to use different methods for handling uncertainty while maintaining the same underlying model (Parsons 1995a). The valuation system that was adopted for handling the non-change data is given in Figure 1—ovals denote variables and boxes denote relations between variables. This network is based upon that in (Smets & Hsia 1990), and expresses the fact that the validity of the structure is a combination of the effect of all constraints, and that the constraints hold by default until they are explicitly represented as failing. Thus, for $C1$ there is a node "$C1$" which is true if $C1$ holds for the structure in question, and false otherwise. The value of this node combined on the node "$C1\&C1A \rightarrow real$" with $p(C1A)$ which is

[‡]In fact, valuation systems are more expressive since they can handle models which include directed cycles, and directed cycles cannot be handled by any form of probabilistic network.

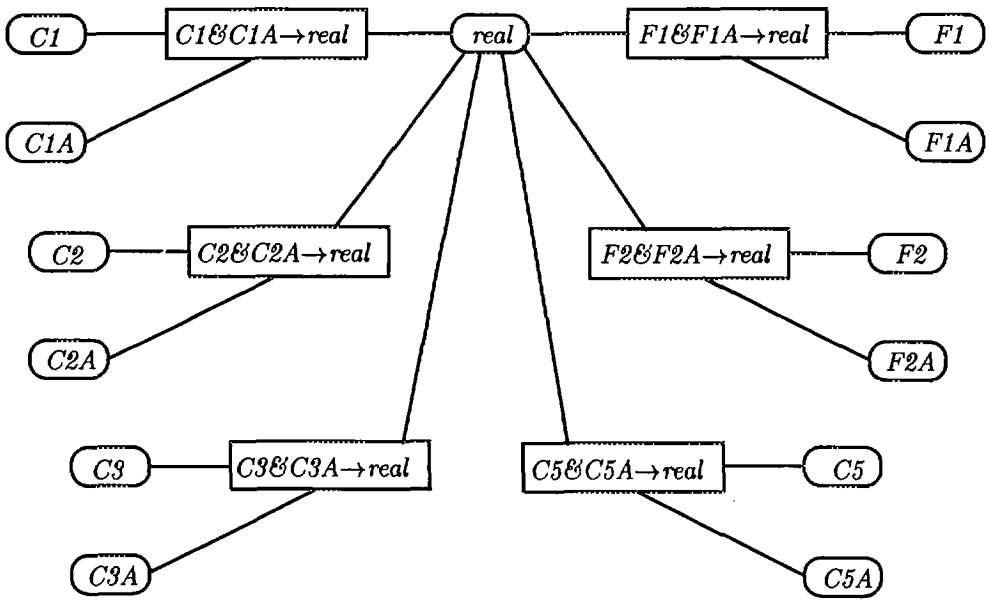

Figure 1: A network relating the effects of constraints to the likelihood that a structure is real.

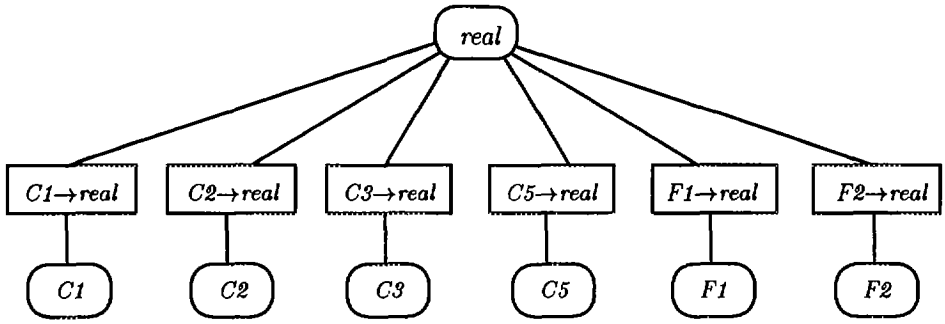

Figure 2: A network for relating changes in information about the applicability of constraints to the changes in likelihood that a structure is real.

the value of node "$C1A$". This is then combined with similar results for other constraints to get an overall measure of the likelihood that the protein is real. This valuation system was then evaluated using Mummu (Parsons 1995b), which handles valuation systems with point and interval values, to get the results presented in the next section. This system could easily be linked to the members of the CBS family of programs, although such work has not been attempted.

It is also possible to construct a network (Figure 2) which relates the qualitative change in probability of occurrence of the individual constraints to the change in the likelihood of a structure being a real protein by using relations "*Cx→real*" which reflect the relevant derivatives. Once again this model can be evaluated using Mummu, and could be linked to members of the CBS family of programs.

Results.

Table 2(a) has point values for each of the constraint probabilities $p(xA)$ that are based on the disambiguated sample of eight nucleotide binding domain proteins. The results of using these in the first valuation system model are given in Figure 3 and the second and fifth columns of Table 4. In the graph each point on the x-axis corresponds to a single set of constraints, with the validity measured up the y-axis, and the leftmost point on the x-axis corresponds to the first set of constraints in Table 4. These results show a range of values from the certainty that a topology will be a real protein when all the constraints hold, to ignorance (a probability of 0.5) when no constraints hold. Note that any structure which conforms to $C1$ is predicted to be certainly real. Table 2(b) has interval values which represent the ambiguity surrounding $F1$ and $F2$

Constraint Set	$p(real)$	$p(real)$	Constraint Set	$p(real)$	$p(real)$
$\{C1, C2, C3, C5, F1, F2\}$	1.0	[1.0 1.0]	$\{C2, C3, F1, F2\}$	0.998	[0.990 0.998]
$\{C1, C2, C3, C5, F1\}$	1.0	[1.0 1.0]	$\{C2, C3, F1\}$	0.988	[0.985 0.998]
$\{C1, C2, C3, C5, F2\}$	1.0	[1.0 1.0]	$\{C2, C3, F2\}$	0.997	[0.955 0.988]
$\{C1, C2, C3, C5\}$	1.0	[1.0 1.0]	$\{C2, C3\}$	0.986	[0.937 0.986]
$\{C1, C3, C5, F1, F2\}$	1.0	[1.0 1.0]	$\{C2, C5, F1, F2\}$	0.996	[0.985 0.997]
$\{C1, C3, C5, F1\}$	1.0	[1.0 1.0]	$\{C2, C5, F1\}$	0.979	[0.978 0.997]
$\{C1, C3, C5, F2\}$	1.0	[1.0 1.0]	$\{C2, C5, F2\}$	0.995	[0.934 0.982]
$\{C1, C3, C5\}$	1.0	[1.0 1.0]	$\{C2, C5\}$	0.975	[0.909 0.979]
$\{C1, C2, C5, F1, F2\}$	1.0	[1.0 1.0]	$\{C3, C5, F1, F2\}$	0.978	[0.970 0.995]
$\{C1, C2, C5, F1\}$	1.0	[1.0 1.0]	$\{C3, C5, F1\}$	0.905	[0.957 0.994]
$\{C1, C2, C5, F2\}$	1.0	[1.0 1.0]	$\{C3, C5, F2\}$	0.974	[0.877 0.965]
$\{C1, C2, C5\}$	1.0	[1.0 1.0]	$\{C3, C5\}$	0.889	[0.833 0.958]
$\{C2, C3, C5, F1, F2\}$	0.999	[0.996 1.0]	$\{C1, F1, F2\}$	1.0	[1.0 1.0]
$\{C2, C3, C5, F1\}$	0.994	[0.994 1.0]	$\{C1, F1\}$	1.0	[1.0 1.0]
$\{C2, C3, C5, F1\}$	0.999	[0.983 0.995]	$\{C1, F2\}$	1.0	[1.0 1.0]
$\{C2, C3, C5\}$	0.993	[0.976 0.995]	$\{C1\}$	1.0	[1.0 1.0]
$\{C1, C2, C3, F1, F2\}$	1.0	[1.0 1.0]	$\{C2, F1, F2\}$	0.991	[0.96 0.993]
$\{C1, C2, C3, F1\}$	1.0	[1.0 1.0]	$\{C2, F1\}$	0.957	[0.944 0.992]
$\{C1, C2, C3, F2\}$	1.0	[1.0 1.0]	$\{C2, F2\}$	0.989	[0.842 0.954]
$\{C1, C2, C3, F1, F2\}$	1.0	[1.0 1.0]	$\{C2\}$	0.950	[0.789 0.945]
$\{C1, C2, F1, F2\}$	1.0	[1.0 1.0]	$\{C3, F1, F2\}$	0.955	[0.923 0.986]
$\{C1, C2, F1\}$	1.0	[1.0 1.0]	$\{C3, F1\}$	0.819	[0.894 0.984]
$\{C1, C2, F2\}$	1.0	[1.0 1.0]	$\{C3, F2\}$	0.947	[0.727 0.911]
$\{C1, C2\}$	1.0	[1.0 1.0]	$\{C3\}$	0.792	[0.651 0.896]
$\{C1, C3, F1, F2\}$	1.0	[1.0 1.0]	$\{C5, F1, F2\}$	0.922	[0.889 0.980]
$\{C1, C3, F1\}$	1.0	[1.0 1.0]	$\{C5, F1\}$	0.715	[0.848 0.976]
$\{C1, C3, F2\}$	1.0	[1.0 1.0]	$\{C5, F2\}$	0.909	[0.64 0.873]
$\{C1, C3\}$	1.0	[1.0 1.0]	$\{C5\}$	0.678	[0.554 0.851]
$\{C1, C5, F1, F2\}$	1.0	[1.0 1.0]	$\{F1, F2\}$	0.849	[0.75 0.947]
$\{C1, C5, F1\}$	1.0	[1.0 1.0]	$\{F1\}$	0.543	[0.677 0.937]
$\{C1, C5, F2\}$	1.0	[1.0 1.0]	$\{F2\}$	0.826	[0.4 0.72]
$\{C1, C5\}$	1.0	[1.0 1.0]	$\{\}$	0.5	[0.318 0.682]

Table 4: Results of the experiment in assessing the validity of sets of constraints

holding. Results using these values are given in the third and sixth columns of Table 4 and Figure 4. In order to represent intervals graphically they have been transformed into point values by replacing them with their mid-points. This transformation may be justified (Parsons 1995b) by an argument based on the principle of maximum entropy. Both sets of results are dominated by the value attached to $C1$, but it is nevertheless clear that as the set of constraints to which a protein conforms is reduced (which is what proceeding along the x-axis represents), its measure of validity decreases, except where a very unreliable constraint is relaxed when there is a sharp increase in validity.

We also have results about the effect of changing the set of constraints to which a structure conforms. Adding constraints one at a time using the network of Figure 2 and establishing the results of the addition using Mummu allows us to evaluate the model that uses qualitative changes. This shows that the addition of all constraints except $F1$ causes $p(real)$ to rise, the latter causing it to fall (Table 5—[+] indicates a rise in $p(real)$, [−] a fall) an outcome which is expected from the data of Table 2. It is less easy to see how these results fit against those in Table 4 and Figures 3–4 since since they are based on a different underlying model. However, both in tabular and graphical form it is clear that removing some constraints, $C1$ for instance, causes the validity of a structure to decrease sharply, while removing less reliable constraints such as $F2$ cause a less dramatic change.

Constraint Added	Change in $p(real)$
$C1$	[+]
$C2$	[+]
$C3$	[+]
$C5$	[+]
$F1$	[−]
$F2$	[+]

Table 5: The results of using the probabilistic qualitative derivatives from Table 3

Discussion.

Unfortunately, there is no obvious "gold standard" (Heckerman 1990) against which to compare the re-

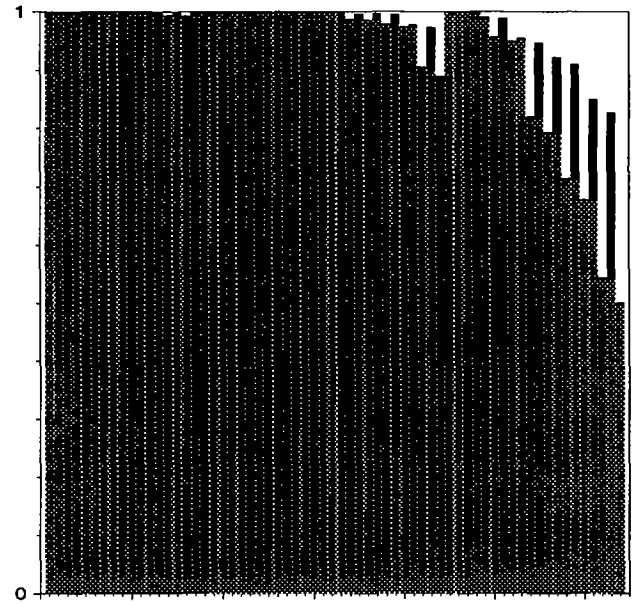

Figure 3: Results based on the point probabilities from Table 2(a)

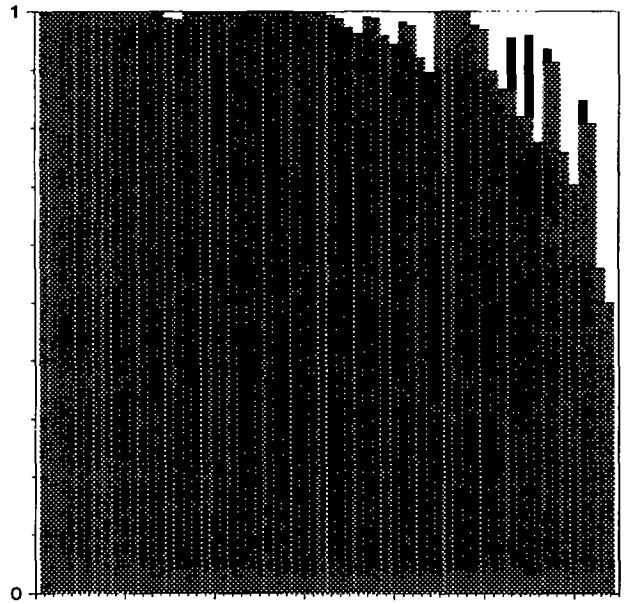

Figure 4: Results based on the interval probabilities from Table 2(b)

sults, so we are forced to use more pragmatic methods of evaluation, and on this basis it is possible to suggest a number of criteria for both choosing between results and deciding whether any of them are useful, while bearing in mind that the intention of this work is to provide a means of focussing attention on a group of likely structures rather than determining the absolute best structure.

For instance the decision about which results are most acceptable will partly depend on which method of dealing with ambiguity is preferred. If the "disambiguated" values are chosen, the results in the second and fifth columns and Figure 3 apply. If an interval representation of the ambiguity is desired, then the third and sixth columns and graph of Figure 4 should be considered.

Thought might also be given to what the results are to be used for, and the decision about whether they are useful made on the basis of which are most useful. In this case it may be of little use having a set of values which contain many identical entries, an argument which suggests that the results might be more useful if they were more disperate since, as they stand, they have a value of 1 for any constraint set containing $C1$. On the other hand this could be acceptable as a clear indication of the necessity of having structures conformant with $C1$.

Another point is that the prediction of protein topology is only a part of the process of establishing structure. Clearly, if a large number of experiments are required in order to reject each possible structure, it would be advantageous to start with the smallest possible set of structures. This suggests considering the number of possible structures associated with various sets of constraints when considering the usefulness of the results. It is possible to determine the number of structures associated with sets of constraints, and the order of the seven sets for which this has been done (Rawlings 1995), based upon the number of possible structures, agrees broadly with the order obtained from our results. This suggests that our results are helpful in this regard, though again more disperate values might make things easier.

Finally, the way in which the topology prediction system is to be used can be considered. If it is intended that the system be used in batch mode to predict a group of structures and their respective validities, then the absolute values given using the first valuation system seem to be the most useful. However, this changes if the system is used interactively, with constraints being added and deleted one by one so that their effect on the structure and the validity can be observed. In this case the use of the qualitative derivatives and the second valuation system seems to be the best alternative since it gives immediate feedback on the change in validity as the constraint set is altered.

Summary.

This paper is not the first to suggest that it would be sensible to model uncertainty in topological constraints. That distinction, to my knowledge, falls to (Clark, Shirazi, & Rawlings 1992). It is also not the

first paper to present work that actually models the uncertainty in the constraints, since, to my knowledge, the first paper to do so was (Clark *et al.* 1994). It does, however, make a number of useful contributions. One of these is the suggestion that valuation systems can be usefully employed in this area. To date valuation systems have had only a fraction of the publicity received by probabilistic networks, yet they are by no means a lesser tool for the modelling of uncertain information, and deserve wider application. In particular, as we have demonstrated in this paper, they are quite appropriate for modelling uncertainty about protein structures, and it is to be hoped that our demonstration convinces others to adopt them in their work.

Another contribution is the demonstration that there are a number of different ways of handling the uncertain information in protein topology prediction, and that all of these may be useful in different situations. In this vein we have shown that there are different ways to handle the ambiguity of the data about the extent to which constraints apply, and that there is merit in simply looking at the way in which the likelihood of a structure being valid changes as well as considering what that likelihood is.

In addition to this the paper has proposed the use of more sophisticated models of handling uncertainty in protein topology prediction than have previously been used. Clark and colleagues (Clark *et al.* 1994) use a simple weighting which records the penalty associated with constraints holding and failing. The weights are obtained in the same way as "pure" probabilities, and combined additively and with the assumption that they are completely independent. Whilst the method presented here also assumes independence, the values we use are more strongly based on objective probability theory, and combined as one would combine probabilities. Furthermore, the use of the valuation system model means that it is easy to extend the approach presented in this paper to use other methods of handling uncertainty (Parsons 1995a), and it is not easy to see how this might be done with simple weights.

Lastly, the rather preliminary nature of this paper should be acknowledged. The work presented here only represents a fraction of the possible work that could be carried out in this area—it really raises many more questions than it answers, and points to a much more significant contribution than the modest effort it records. For instance, the models we have adopted are very simple, and could be greatly refined by considering the dependencies between the constraints, by obtaining more data on the applicability of the constraints, or by using different methods for handling the rather imperfect data that is available. Given time such refinements could be incorporated into the methods outlined in this paper, and it is hoped that such work will be undertaken in the future. In addition, it would be extremely useful to find some means of establishing a "gold standard" against which the results can be judged, and again it is hoped that such work can be undertaken.

Acknowledgments.

I am endebted to Dominic Clark for his help in understanding both the domain and the constraint-based approach of CBS1, and to Chris Rawlings for pointing out problems with some of the data that I was originally using.

References

Allison, L. 1993. Methods for dealing with error and uncertainty in molecular biology computations and databases. In *Proceedings of the 22nd Hawaiian International Conference on System Sciences*, 704. Los Alamitos, California: IEEE Computer Society Press.

Altman, R. B., and Jardetzky, O. 1989. Heuristic refinement method for determination of solution structure of proteins from nuclear magnetic resonance. *Methods in Enzymology* 177:218–246.

Altman, R. B. 1994a. Constraint satisfaction techniques for modelling large complexes: application to the central domain of 16S ribosomal RNA. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 10–18. Menlo Park, California: AAAI Press.

Altman, R. B. 1994b. Probabilistic structure calculations: a three dimensional tRNA structure from sequence correlation data. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 12–20. Menlo Park, California: AAAI Press.

Asai, K.; Hayamizu, S.; and Onizuka, K. 1993. HMM with protein structure grammar. In *Proceedings of the 22nd Hawaiian International Conference on System Sciences*, 783–791. Los Alamitos, California: IEEE Computer Society Press.

Brown, M.; Hughey, R.; Krogh, A.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1994. Using Dirchlet mixture priors to derive hidden markov models for protein families. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 47–55. Menlo Park, California: AAAI Press.

Clark, D. A.; Rawlings, C. J.; Shirazi, J.; Veron, A.; and Reeve, M. 1994. Protein topology predic-

tion through parallel constraint logic programming. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 83–91. Menlo Park, California: AAAI Press.

Clark, D. A.; Rawlings, C. J.; and Doursenot, S. 1994. Genetic map construction with constraints. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 78–86. Menlo Park, California: AAAI Press.

Clark, D. A.; Shirazi, J.; and Rawlings, C. J. 1992. Protein topology prediction through constraint-based search and the evaluation of topological folding rules. *Protein Engineering* 4:751–760.

Cohen, F. E., and Kuntz, I. D. 1987. Prediction of the three dimensional structure of human growth hormone. *Proteins: Structure, Function and Genetics* 2:162–167.

Delcher, A. L.; Kasif, S.; Goldberg, H. R.; and Hsu, W. H. 1993. Probabilistic prediction of protein secondary structure using causal networks. In *Proceedings of the 11th National Conference on Artificial Intelligence*, 316–321. Menlo Park, California: AAAI Press.

Delcher, A. L.; Kasif, S.; Goldberg, H. R.; and Hsu, W. H. 1994. Protein structure modelling using probabilistic networks. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 109–117. Menlo Park, California: AAAI Press.

Haussler, D.; Krogh, A.; Mian, I. S.; and Sjölander, K. 1993. Protein modelling using Hidden Markov Models: Analysis of globins. In *Proceedings of the 22nd Hawaiian International Conference on System Sciences*, 792–802. Los Alamitos, California: IEEE Computer Society Press.

Heckerman, D. E., and Shwe, M. 1993. Diagnosis of multiple faults: a sensitivity analysis. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, 80–87. San Mateo, California: Morgan Kaufmann.

Heckerman, D. E. 1990. An empirical comparison of three inference methods. In Shachter, R. D.; Levitt, T. S.; Kanal, L. N.; and Lemmer, J. F., eds., *Uncertainty in Artificial Intelligence 4*. Amsterdam: Elsevier. 283–302.

King, R. D.; Clark, D. A.; Shirazi, J.; and Sternberg, M. J. E. 1994. Inductive logic programming used to discover topological constraints in protein structures. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 219–226. Menlo Park, California: AAAI Press.

MacKenzie, T.; Platt, D.; and Dix, T. 1993. Modelling errors in restriction mapping. In *Proceedings of the 22nd Hawaiian International Conference on System Sciences*, 613–619. Los Alamitos, California: IEEE Computer Society Press.

Parsons, S. 1995a. Hybrid models of uncertainty in protein topology prediction. *Applied Artificial Intelligence* 9:335–351.

Parsons, S. 1995b. *Qualitative approaches to reasoning under uncertainty*. Cambridge, Massachusetts: MIT Press.

Rawlings, C. J. 1995. Personal communication.

Saffiotti, A.; Parsons, S.; and Umkehrer, E. 1994. A case study in comparing uncertainty management techniques. *Microcomputers in Civil Engineering: Special Issue on Uncertainty in Expert Systems* 9:367–380.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton, New Jersey: Princeton University Press.

Shenoy, P. P. 1991. A valuation-based language for expert systems. *International Journal of Approximate Reasoning* 3:383–411.

Smets, P., and Hsia, Y.-T. 1990. Default reasoning and the transferable belief model. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 529–537. Mountain View, California: Association for Uncertainty in AI.

Spiegelhalter, D. 1990. Fast algorithms for probabilistic reasoning in influence diagrams, with applications in genetics and expert systems. In Oliver, R. M., and Smith, J. Q., eds., *Influence Diagrams, Belief Nets and Decision Analysis*. New York: John Wiley & Sons Ltd. 361–383.

Szolovits, P. 1992. Compilation for fast calculation over pedigrees. *Cytogenetics and Cell Genetics* 59:136–138.

Taylor, W. R., and Green, N. M. 1989. The predicted secondary structure of the nucleotide binding sites of six cation-transporting ATPases leads to a probable tertiary fold. *European Journal of Biochemistry* 179:241–248.

Zadeh, L. A. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1:1–28.