Using qualitative uncertainty in protein topology prediction

Simon Parsons

[1] Advanced Computation Laboratory, Imperial Cancer Research Fund, P.O. Box 123, Lincoln's Inn Fields, London WC2A 3PX, United Kingdom.
[2] Department of Electronic Engineering, Queen Mary and Westfield College, Mile End Road, London, E1 4NS, United Kingdom.

Abstract. The prediction of protein structure is an important problem in molecular biology. It is also a difficult problem since the available data are incomplete and uncertain. This paper describes models for the prediction of a particular level of protein structure, known as the topology, which handle uncertainty in a qualitative fashion.

1 Introduction

Proteins are large biological macromolecules that form the main components of living organisms and control most of the crucial processes in within them. The function of a particular protein is determined by the chemical interactions at its surface, and these are related to its three dimensional structure. Thus knowledge of protein structure is important. The structure of proteins can be described at various levels of detail from the primary structure, which consists of a list of the amino acids that make up the protein, through the secondary structure, which is a description of the way that the amino acids are grouped together into substructures such as β-strands and α-helices, to the tertiary structure, which is the set of three dimensional co-ordinates of every atom in the protein. Protein topology is an intermediate level somewhere between secondary and tertiary structure which specifies how the substructures are arranged.

Now, knowledge of three dimensional protein structure is sparse so that while the primary structures for many tens of thousands of proteins are known, only some hundreds of distinct proteins have had their three dimensional structure determined. This discrepancy motivates much research into determining protein structure including the use of computational techniques.

2 Protein Topology Prediction

The prediction of protein topology is interesting because the topology can be used to guide the choice of experiments to confirm protein structure. A major difficulty in this prediction is that a vast number of possible topologies can be hypothesized from a single secondary structure prediction, and one means to tackle this problem is to identify and apply constraints based upon analyses

of known protein structures. For instance, for α/β sheets [1, 13] (which are topological structures combining α-helices and β-strands):

- C1. For parallel pairs of β-strands, β-α-β and β-coil[3]-β connections are right handed.
- C2. The initial β-strand is not an edge strand in the sheet.
- C3. Only one change in winding direction occurs.
- C5. All strands lie parallel in the β-sheet.
- F1. Strands are ordered in the sheet by hydrophobicity, with the most hydrophobic[4] strands central.
- F2. Parallel β-coil-β connections contain at least 10 amino acids.

Because these constraints are derived from aggregate properties of a collection of proteins, they do not apply to all proteins. When Shirazi et al. [12] assessed the validity of $C1$, $C2$, $C3$, $C5$ and $F2$ by checking them against 33 α/β sheet proteins, they found that only one protein satisfied all the constraints. Their results, reproduced in Table 1, show that 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 [7]. This paper explores an alternative method which fits in well with the constraint-satisfaction approach to protein topology prediction reported by Clark et al. [1].

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. Thus following each step a structure can either conform to the same set of constraints as before, or to some subset or superset of it. 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 that structure being correct, then the effect of the new knowledge may be

Protein ID	Constraints Violated	Protein ID	Constraints Violated	Protein ID	Constraints Violated
p1aat	C2 C5	p1ts1	C2 C5	p1ppd	C2 C5
p1bp2	C2 C5	p1ubq	C3 C5	p1rn3	C2 C5
p1cac	C2 C3 C5	p2b5c	C2 C3 C5	p1sbt	C1
p1cpb	C2 C3 C5	p2cab	C2 C3 C5	p1sn3	C2 C5
p1crn	C2 C5	p2cdv	C2 C5	p1srx	C2 C3 C5
p1cts	C2 C5	p2cts	C2 C5	p5cpa	C2 C3 C5
p1ctx	C5	p2lzm	C5	p3pgm	C5
p1hip	C2 C5	p2ssi	C5	p4cts	C2 C5
p1nxb	C3 C5	p3bp2	C2 C5	p4dfr	C3 C5 F2
p1ovo	C5	p3cts	C2 C5	p4fxn	
p1p2p	C2 C5	p3dfr	C3 C5	p4pti	C2 C5

Table 1. The results of checking constraints against 33 α/β sheet proteins.

[3] A protein has coil structure where it is neither a β-strand nor an α-helix.
[4] Lacking an affinity for water.

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.

Now, information about changes in the validity of a structure being correct with changes in evidence about which constraints it conforms to is exactly the kind of information that is handled by our qualitative approach to propagating uncertainty [8], and methods based upon this approach are what we consider here. In the tradition of experimental investigations of how to model uncertainty in a given problem [3, 4, 7, 10] 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, and some of these are discussed in [6].

3 Single formalism approaches

The data in Table 1 may 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 $C1$:

$$p(C1\,|\,real) = \frac{\text{Number of proteins for which } C1 \text{ holds}}{\text{Total number of proteins}} = \frac{32}{33}$$

We have no information about the proportion of proteins for which $C1$ holds yet which are not real, so we cannot establish $p(C1\,|\,\neg real)$ in the same way. Instead, we must employ the principle of maximum entropy to conclude that $p(C1\,|\,\neg real) = 0.5$. From [8] we learn that these values are sufficient to establish the relationship between $p(C1)$ and $p(real)$ as being that $\frac{dp(C1)}{dp(real)} = [+]$, so that as $p(real)$ increases, so does $p(C1)$. This information, in turn [8], tells us that $\frac{dp(real)}{dp(C1)} = [+]$, allowing us to establish how $p(real)$ changes when we have information about $C1$ holding. Using the data about other constraints, we get Table 2. Note that $\frac{dp(real)}{dp(C2)} = [-]$ indicates that as $p(C2)$ increases, $p(real)$ decreases.

Constraint (x)	Cases of constraint failure	$\frac{dp(real)}{dp(x)}$	Change in $p(real)$ on adding the constraint
$C1$	1	[+]	[+]
$C2$	23	[−]	[−]
$C3$	10	[+]	[+]
$C5$	31	[−]	[−]
$F2$	1	[+]	[+]

Table 2. The probabilistic qualitative derivatives and their effects

It is possible to construct a valuation system model [11] which allows us to combine the effects of the various constraints. A suitable network is given in Fig 1—ovals denote variables, and boxes denote relations between variables. The propagation of qualitative values in this network may be carried out by the Mummu system [8], and using Mummu we can establish that the addition of $C1$, $C3$ and $F2$ causes $p(real)$ to rise, while the addition of $C2$ and $C5$ cause it to fall (Table 2).

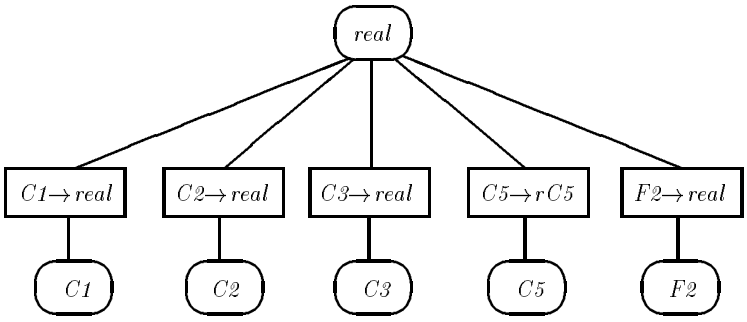

Fig. 1. A network for propagating qualitative changes

It is also possible to model the constraints using possibility theory. If a structure conforms to a constraint, then it is entirely possible that the structure is correct. However, if a structure fails to conform to a constraint then it becomes less possible that the structure is correct. Indeed, the possibility of the structure being a protein falls to a figure that reflects the proportion of naturally occurring proteins that do not conform to the constraint. So, considering the data in Table 1, we have:

$$\Pi(C1\,|\,real) = \frac{\text{Number of proteins for which } C1 \text{ does not hold}}{\text{Total number of proteins}} = \frac{1}{33}$$

Since we have no information about proteins which are not real, we know nothing about $\Pi(C1\,|\,\neg real)$ and $\Pi(\neg C1\,|\,\neg real)$, and so set them both to 1 by the principle of minimum specificity [2]. These values, along with $\Pi(real) = \Pi(\neg real) = 1$ (again by the principle of minimum specificity) allow us to establish derivatives that define the relationship between $\Pi(real)$ and $\Pi(C1)$ [8] to be $\frac{d\Pi(C1)}{d\Pi(real)} = [0]$, $\frac{d\Pi(\neg C1)}{d\Pi(real)} = [\downarrow]$, $\frac{d\Pi(C1)}{d\Pi(\neg real)} = [0]$ and $\frac{d\Pi(\neg C1)}{d\Pi(\neg real)} = [0]$, meaning that $\Pi(\neg C1)$ may decrease when $\Pi(real)$ decreases, whilst it is independent of $\Pi(\neg real)$, and $\Pi(C1)$ is independent of $\Pi(real)$ and $\Pi(\neg real)$. From these values it is possible [8] to determine that $\frac{d\Pi(real)}{d\Pi(\neg C1)} = [\downarrow]$ with the other derivatives concerning $C1$ all being zero. Similar reasoning about the other constraints gives Table 3. When these derivatives are used with the network in Fig 1, and the effects of the application of individual constraints are propagated using Mummu, the results of the last column of Table 3 are generated. These results are rather different from

Constraint (x)	$\frac{d\Pi(real)}{d\Pi(x)}$	$\frac{d\Pi(real)}{d\Pi(\neg x)}$	$\frac{d\Pi(\neg real)}{d\Pi(x)}$	$\frac{d\Pi(\neg real)}{d\Pi(\neg x)}$	Change in $\Pi(real)$ on removing the constraint
$C1$	[↓]	[0]	[0]	[0]	[−]
$C2$	[↓]	[0]	[0]	[0]	[−]
$C3$	[↓]	[0]	[0]	[0]	[−]
$C5$	[↓]	[0]	[0]	[0]	[−]
$F2$	[↓]	[0]	[0]	[0]	[−]

Table 3. The possibilistic qualitative derivatives and their effects

those generated by the probabilistic modelling given above since they predict a change in possibility when a constraint is violated rather than a change in probability when a constraint is conformed to. At first sight it might appear that those constraints that, when added, cause a decrease in probability (that is $C2$ and $C5$), should, when removed, cause an increase in possibility. However, on reflection, this is seen not to be the case. Since, under our interpretation, violation of a constraint simply means that the possibility of a structure falls to reflect the proportion of structures that violate the constraint, when $C2$ and $C5$ are violated, the fall in possibility still occurs—it is just smaller than for other constraints.

4 Integrated approaches

It is also possible to integrate different representations of uncertainty using qualitative changes [8, 9], and this enables us to model the protein topology prediction problem in a slightly different way. There is another set of data about the applicability of the constraints [1, 7], which identifies some ambiguity in the data. This arises because there were a number of alternative structures for some of the proteins that were tested, and the constraints applied to some of these structures but not to others. In particular, $F1$ was found to hold for 1 of the 8 proteins tested, be violated for 5 of the proteins, and be ambiguous for 2, while $F2$ held for 6, was violated for 1, and was ambiguous for 1.

One way of modelling this ambiguity is to use Dempster-Shafer theory, and if the basic probability assignments that follow from the data given above are taken and interpreted as conditional beliefs, in the same way as the probabilistic data has previously been interpreted, then $bel(\{F1\} \mid \{real\}) = 0.125$, $bel(\{\neg F1\} \mid \{real\}) = 0.625$, $bel(\{F1, \neg F1\} \mid \{real\}) = 0.25$. Since there is no data about proteins that are not real we employ the Dempster-Shafer model of ignorance to get $bel(\{F1\} \mid \{\neg real\}) = 0$, $bel(\{\neg F1\} \mid \{\neg real\}) = 0$, $bel(\{F1, \neg F1\} \mid \{\neg real\}) = 1$, $bel(\{F1\} \mid \{real, \neg real\}) = 0$, $bel(\{\neg F1\} \mid \{real, \neg real\}) = 0$ and $bel(\{F1, \neg F1\} \mid \{real, \neg real\}) = 1$. These values tell us [8] that $\frac{dbel(\{F1\})}{dbel(\{real\})} = [+]$, $\frac{dbel(\{\neg F1\})}{dbel(\{real\})} = [+]$ and $\frac{dbel(\{F1, \neg F1\})}{dbel(\{real\})} = [-]$ and these may be transformed [8] to give $\frac{dbel(\{real\})}{dbel(\{F1\})} = [+]$, $\frac{dbel(\{real\})}{dbel(\{\neg F1\})} = [+]$ and $\frac{dbel(\{real\})}{dbel(\{F1, \neg F1\})} = [-]$. All other derivatives relating $F1$ and $real$ have value [0]. Repeating this procedure for $F2$

Constraint (x)	$\frac{dbel(\{real\})}{dbel(\{x\})}$	$\frac{dbel(\{real\})}{dbel(\{\neg x\})}$	$\frac{dbel(\{\neg real\})}{dbel(\{x\})}$	$\frac{dbel(\{\neg real\})}{dbel(\{\neg x\})}$	$\frac{dbel(\{real\})}{dbel(\{x,\neg x\})}$	$\frac{dbel(\{\neg real\})}{dbel(\{x,\neg x\})}$
$F1$	$[+]$	$[+]$	$[0]$	$[0]$	$[-]$	$[0]$
$F2$	$[+]$	$[0]$	$[0]$	$[0]$	$[-]$	$[0]$

Table 4. The Dempster-Shafer qualitative derivatives

gives the derivatives of Table 4.

These values may be used in conjunction with the probabilistic ones given above in the network of Fig 2. This network is simply that of Fig 1 extended to include the dependency of *real* on $F1$. The relationships between $C1$, $C2$, $C3$, $C5$ and *real* are determined using qualitative probabilities, while those between $F1$, $F2$ and *real* are determined using qualitative beliefs. The approach using qualitative changes that we are employing allows the combined use of different

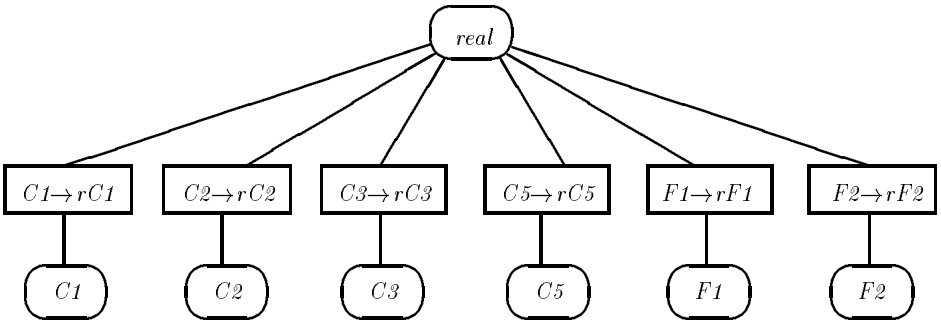

Fig. 2. A second network for propagating qualitative changes

formalisms together, simply translating a value of $[+]$ (a definite increase) in belief functions to $[0,+]$ (a possible increase) in probability, and $[-]$ into $[0,-]$, and giving the overall change at *real* as a qualitative probability. As before it is possible to consider changes in the value of *real* when new evidence is obtained

Constraint Added	Change in probability of *real*
$C1$	$[+]$
$C2$	$[-]$
$C3$	$[+]$
$C5$	$[-]$
$F1$	$[+,0]$
$F2$	$[+,0]$

Table 5. The results of using the probabilistic and Dempster-Shafer qualitative derivatives

Constraint Added	Change in possibility of $real$
$C1$	$[0]$
$C2$	$[0]$
$C3$	$[0]$
$C5$	$[0]$
$F1$	$[+, 0]$
$F2$	$[+, 0]$

Constraint Violated	Change in possibility of $real$
$C1$	$[-]$
$C2$	$[-]$
$C3$	$[-]$
$C5$	$[-]$
$F1$	$[-, 0]$
$F2$	$[-, 0]$

Table 6. The results of using the possibilistic and Dempster-Shafer qualitative derivatives

about a constraint holding, since Mummu implements the integration of changes in value discussed in [8, 9]. The results of applying Mummu are given in Table 5.

It is also possible to integrate possibility and belief values using the same network. Belief changes concerning $F1$ and $F2$ are propagated using the derivatives in Table 4, and changes in possibilities concerning $C1$, $C2$, $C3$ and $C5$ are propagated using the derivatives of Table 3. Translation from beliefs to possibilities are carried out in the same way as from beliefs to probabilities, and the overall change in $real$ is given as a qualitative possibility. Since changes in possibility of the leaf nodes of the network in only occur when constraints are violated, both the addition and violation of constraints is considered. This set-up generates the results of Table 6.

Thus the use of both the single and combined formalism approaches make it possible to establish the change in validity of protein structure as components are added. The qualitative information that is provided is sufficient to assess how valid the addition is, and thus is sufficient to guide the addition of components during the constraint-based search.

5 Discussion

Unfortunately there is no obvious "gold standard" [3] against which to compare the results so that it is not possible to prove that they are helpful. However, it is possible to make several arguments for their being worth having and for the modelling experiment having been worthwhile. Firstly, it is a demonstration that purely qualitative methods for handling uncertainty can be useful. Thus it provides a useful counterpart to [5], which showed that qualitative probability could be usefully used in a diagnosis problem. Secondly it extends the comparative study of the use of differing uncertainty handling techniques [3, 4, 7, 10] to cover a new problem—that of modelling the impact of changing constraints in protein topology prediction. This problem contains a number of different types of uncertainty, and the fact that different models seem appropriate from different points of view provides empirical evidence for the validity of work on the different models. In addition, since no model seems to naturally model every aspect of the uncertainty, the protein topology problem provides motivation for working

on using the different models in combination in the same problem. Further to this motivation, this paper, as is the case with the companion paper [7], has suggested some means of combining different methods within one problem, and, using results generated using the implementation of this work in the Mummu system, has illustrated the use of combinations of formalisms in solving a real problem. Thus the paper has provided some empirical demonstration that using combinations of formalisms is both feasible and useful.

References

1. 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.
2. D. Dubois and H. Prade 1991 Fuzzy sets in approximate reasoning, Part1: inference with possibility distributions, *Fuzzy sets and systems*, **40**:143–202.
3. Heckerman, D. E. 1990. An empirical comparison of three inference methods. In *Uncertainty in Artificial Intelligence 4*, (R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, eds.), Elsevier, Amsterdam.
4. Heckerman, D. E., and Shwe, M. 1993. Diagnosis of multiple faults: a sensitivity analysis, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, Washington D. C.
5. Henrion, M., Provan, G., del Favero, B., and Sanders, G. 1994. An experimental comparison of diagnostic performance using infinitesimal and numerical bayesian belief networks, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, Seattle.
6. Parsons, S. 1995. Softening constraints in constraint-based protein topology prediction, *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*, Cambridge, UK.
7. Parsons, S. 1995. Hybrid models of uncertainty in protein topology prediction, *Applied Artificial Intelligence*, **9**:335–351.
8. Parsons, S. 1993. Qualitative methods for reasoning under uncertainty, PhD Thesis, Queen Mary and Westfield College, London (to be published by MIT Press).
9. Parsons, S. and Saffiotti, A. 1993. Integrating uncertainty handling techniques in distributed artificial intelligence, in: *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, (M. Clarke, R, Kruse and S. Moral, eds.), Springer Verlag.
10. 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.
11. Shenoy, P. P. 1991. A valuation-based language for expert systems, *International Journal of Approximate Reasoning*, **3**:383–411.
12. Shirazi, J., Clark, D. A. and Rawlings, C. J. 1990. Constraint-based reasoning in molecular biology: predicting protein topology from secondary structure and topological constraints, BCU/ICRF Technical Report.
13. 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.

This article was processed using the LaTeX macro package with LLNCS style