Paper presented at the UNICOM Seminar and Tutorial on Causal Models and Statistical Learning, London, March 1997.

CAUSALITY, STATISTICAL LEARNING AND MULTIPLE ALIGNMENT

J Gerard Wolff

March 1997

School of Electronic Engineering and Computer Systems, University of Wales, Dean Street, Bangor, LL57 1UT, UK.

1 INTRODUCTION

Recent research [10-19] has developed the 'SP' conjecture that: *all kinds of computing and formal reasoning may usefully be understood as information compression by multiple alignment, unification and search*, where *unification* means a simple merging of patterns that match each other[1] and *multiple alignment* has a meaning which is close to its meaning in bio-informatics.[2]

It appears that this framework of ideas has potential to integrate and simplify a range of concepts in computing and cognition including unsupervised learning [13-15], information retrieval [16], deductive and probabilistic inference [11, 13], parsing [10] and others [12].

The purpose of this paper is to consider whether or how "information compression by multiple alignment, unification and search" may illuminate concepts of causality. For convenience, the expression just quoted will be abbreviated as 'ICMAUS' and "information compression" will be shortened to 'IC'.

In what follows, I shall first define the meanings of 'cause' and 'effect' as they will be understood in this paper. Then I will review the ICMAUS concepts as they have been developed in this programme of research, including a section on the quantification of IC in the context of ICMAUS. The following two sections will consider, first, how the learning of causal connections may be understood in terms of ICMAUS and, secondly, how causal

1. Although there is some risk of confusion with the related but more complicated concept of 'unification' in logic, the word has been adopted here because no other word seems to fit the intended meaning so well.

2. Information compression may be interpreted as a process of removing unnecessary (redundant) complexity in information - and thus maximising 'simplicity' - whilst preserving as much as possible of its non-redundant descriptive 'power'. Hence the name 'SP'.

inferences may be understood in these terms.

2 WHAT IS A CAUSE?

In this paper, we shall sidestep any lengthy discussion of the usage of words like ‘cause’ and ‘effect’ by adopting the meanings proposed by John Stuart Mill which appear to capture much of the substance of the concepts of causality as they are commonly understood [6].

2.0.1 Cause as a set of sufficient conditions for an effect to occur

In this view, a ‘cause’ (C) for some ‘effect’ (E) is the *set of sufficient conditions* for E to occur, where E is some kind of event or set of events. If all the conditions in C obtain, then E is certain to occur (or as certain as we can be in this uncertain world). Part of this concept of causality (and others) is that C occurs before E or, possibly, at the same time as E, but C *never* occurs after E.

The number of conditions which are sufficient to ensure that a given effect occurs is normally quite large. For example, the sufficient conditions for a car to go includes the presence of most of the parts in the engine, transmission and wheels - in the correct configuration and in working order. It also includes the presence of suitable fuel in the tank and fuel line, sufficient oxygen in the air, the ignition switched on, starter motor turning the engine over and, after the engine starts, gears engaged and the clutch released.

2.0.2 Economy in communication

Given the length of a typical list of causal factors, it is common for people to save breath or ink by identify the cause of an effect as the part or parts of the list of conditions which differ from expectations, assuming that the other parts are already known [6]. Thus, if a car which was not running is made to go we might identify a new battery or new spark plugs as the causal factor without bothering to mention all the other conditions which must obtain for the car to go.

In this paper we shall assume that this aspect of the way people talk about causes is a feature of how we save effort in communication which does not invalidate the underlying concept of a cause as the set of sufficient conditions for the corresponding effect to occur.

2.0.3 Asymmetry in the relations between cause and effect

One other feature of causality deserves a mention here: there is nothing in the definition of a cause which we have adopted to suggest that any given effect must be the result of a single set of causal conditions. It often happens that a given effect may be the result of two or more *alternative* sets of causal conditions. We can, for example, become hot and flushed because we have ’flu, or because we have been running fast or because we are feeling embarrassed, and so on. The sets of causal conditions in cases like these are not necessarily disjoint but they are, nevertheless, distinct alternatives in the way a given effect or set of effects may be caused. In short, a given set of causal conditions must be *sufficient* for the corresponding effect to occur but it is often not a *necessary* precursor or accompaniment of the effect.

Given that our adopted definition of cause implies a single effect or set of effects for any set of causal conditions, there is an asymmetry in the relation between cause and effect:

there may be alternative causes for an effect but not the other way round. We shall return to this point in Section 6 where we consider how deductive and abductive inferences about causes and effects may be drawn.

3 INFORMATION COMPRESSION BY MULTIPLE ALIGNMENT, UNIFICATION AND SEARCH (ICMAUS)

This section summarises the main concepts in the ICMAUS framework as it has been developed to date, except for the quantification of compression which is described in Section 4.

Since the main focus of this paper is on the way in which causality may be modelled by ICMAUS, this section omits many details and does not include any lengthy discussion or justification of the concepts which are described. A fuller account, with more discussion, may be found in [10, 14, 15].

3.1 The SP52 model

The first of the articles just referenced [10] describes a computer model, SP51, which embodies some but not all the concepts described in this section, as will be indicated at appropriate points below. In broad terms, the SP51 model is intended as a model of parsing or recognition of new patterns in terms of patterns which are already known, using the principles of ICMAUS, but without any learning of new patterns. At some point fairly soon it is intended that the model will be generalised so that it can learn new patterns in accordance with the principles of ICMAUS, retaining its capabilities for parsing and recognition.

Work is now nearing completion to generalise the SP51 model so that it can make the kinds of calculations of probability which are described later. The descriptions which follow will be of the generalised version, SP52.

3.2 Basic concepts

3.2.1 Patterns, sequences and symbols

In this research, a *pattern* is understood as an array of symbols (defined below) in one, two or more dimensions. For the sake of simplicity and clarity in this paper, one-dimensional patterns will be the focus of attention but it should be born in mind that the concepts may be generalised to higher dimensions.[3]

A *sub-string* of a (one-dimensional) pattern is one or more symbols within the pattern where each symbol is contiguous with its neighbour or neighbours (if any) within the pattern. A *sub-sequence* of a (one-dimensional) pattern is one or more symbols within the pattern where each symbol need not be contiguous with any neighbours it has within the pattern.

3. Likewise, it should also be born in mind that the digital analysis as described in this paper may be generalised to analogue patterns since any analogue pattern may be digitised with any desired level of precision.

Every sub-string is also a sub-sequence. Sub-strings and sub-sequences may themselves be described as patterns. The length of any sub-string or sub-sequence may range from 1 up to and including the length of the pattern in which it appears.

A *symbol* is some kind of atomic mark which is the smallest part of a pattern which can participate in matching: a symbol may be compared (matched) only with another symbol and the result of matching is either that the two symbols are the same (a *hit*) or that they are different. No other result is permitted.

In this paper, and others in this programme of research, the convention has been adopted that a symbol is represented by any string of one or more characters bounded by spaces.

3.2.2 Alignment

In the case of one-dimensional patterns, an *alignment* is a two-dimensional array:

- Where each row of the array records the symbols in one pattern, no more than one symbol to each cell of the array, with the symbols in the correct order but not necessarily occupying contiguous cells of the array.

- Where sets of two or more symbols which match each other (with each symbol in a different row of the alignment from any other in the set) are arranged in vertical *columns* and marked with linking lines or other method of showing that they match. Columns in the array may include symbols which do not match each other but in cases like this there would be no links.

By contrast with the concept of alignment as it is understood in bio-informatics, any one pattern may appear one, two or more times in an alignment as understood here. This means that any pattern may be aligned with itself. For this to be meaningful, it is obviously necessary to apply the restriction that any one symbol must never be matched against itself and must never appear in the same column as itself.[4]

3.2.3 The nature of redundancy

The key to lossless or lossy IC is the extraction of *redundancy* from information. Before we proceed, a few words are in order about the concept of redundancy which has been adopted in this research.

A working hypothesis in this programme of research is that *all* kinds of redundancy may ultimately be understood as *repetition of patterns*. The patterns which are repeated may be whole patterns or they may be sub-strings within a larger pattern or sub-sequences within a larger pattern. Whether or not the hypothesis turns out to be true, it is clear that much redundancy can be understood in these terms.

A key point is that repetition of patterns, in itself, is not sufficient to signal redundancy: any sequence of random numbers (which, if is truly random, contains no redundancy) may be

---

4. In bio-informatics, an alignment may contain two or more sequences which match each other perfectly but, by contrast with the present framework, they are regarded as sequences with distinct identities, not multiple appearances of a *single* sequence. Hence the restriction just mentioned does not apply.

seen to contain repeating patterns. The key point is that redundancy exists if *a pattern repeats more frequently than other patterns of the same size* or if *a pattern is longer than other patterns of the same frequency*.

Many approaches to IC emphasise frequency of occurrence as a guide to redundancy and pay less attention to the equally important fact that redundancy may manifest itself in the form of patterns with low frequency of occurrence (perhaps as low as 2) but relatively large size.

The roles of frequency and sizes of patterns in the ICMAUS framework are described briefly in Section 4, below.

3.2.4 Unification and search

As previously noted, *unification* in this paper means a simple merging of matching patterns or parts of patterns. It provides the basic mechanism by which the compression of an alignment may be achieved (more below).

Given the overall goal of finding alignments which yield as much compression as can be achieved with the available resources, and given that, normally, the number of alternative alignments for a given set of patterns is astronomically large, a process of 'heuristic' *search* is required which explores the space of possible alignments in stages, using some kind of measure of potential compression of alignments to narrow the search space at each stage, thus avoiding the excessive cost of any kind of exhaustive search. Techniques of this kind include 'hill climbing' ('descent'), 'beam search', 'genetic algorithms', 'simulated annealing', and others.

3.3 The 'Corpus', Minimum Message Length encoding (MML), incremental search, 'Old' and 'New'

In the ICMAUS framework, the set of patterns which are the focus of interest is termed the *Corpus*, a term borrowed from theoretical linguistics where it means the entire body of linguistic information from which a grammar may be derived.

In accordance with the principle of Minimum Message Length encoding (MML) (see, for example, [7, 9]), the goal in the ICMAUS framework is to compress the Corpus so that the compressed version of the corpus includes both the coding system created for the purpose of compression and the information which has been encoded in terms of that system.

In terms of practicalities, there is a problem: in most cognitive systems (e.g., human cognition), the Corpus is constantly growing as new information arrives. It would be rather inefficient if were to re-process the entire Corpus every time new information arrives because we would often be re-processing information which had been compressed in earlier cycles.

For this reason, an incremental approach to compression has been adopted. The Corpus is regarded as being divided into two parts: 'Old' information which is considered to be compressed already as much as is practicable, or which, for reasons connected with the application in hand, does not need to be compressed; and 'New' information which is not yet compressed.

The ICMAUS process attempts to compress New, section by section, by ‘encoding’ each section as far as possible in terms of patterns in Old (see below) and then adding it to Old, either in compressed form (when encoding has been possible) or in its unencoded form (when encoding has not been possible).[5] In this way, the system avoids constant re-processing of the information in Old.

‘Parsing’ or ‘recognition’ may be deemed to occur when patterns in New can be encoded in terms of patterns already stored in Old and which have already been assigned ‘code’ symbols and ‘boundary’ symbols (see next section).

‘Learning’ may be deemed to occur when patterns from New cannot be coded efficiently in terms of patterns already stored in Old and are therefore added to Old in uncompressed form. The concept of ‘learning’ may also be deemed to include the process of adding new ‘code’ symbols to patterns which have been stored in Old (see next section).

Each alignment is evaluated in terms of the compression that may be achieved if New (or the relevant section of New) were to be encoded in terms of the patterns from Old which appear in the alignment.

3.4 Encoding at one level

The basic mode of operation of the ICMAUS process is the basis of all the ‘standard’ methods of IC [8]: find repeated instances of some string of symbols and replace each instance with a relatively short ‘code’.

Where a sub-string of New matches a sub-string in Old, that section may be unified with the matching symbols in Old - and thus, in effect, deleted from New. The position of the section within New may then be marked by some relatively short ‘code’:

- The code symbols may be newly minted if there is nothing already assigned to the relevant pattern in Old, or
- The code symbols installed in New may be a copies of symbols already attached to the relevant pattern in Old.

An example of the first kind, which echoes the use of this technique in the text of this paper, is shown in Fig. 1.

Fig. 1. An alignment and unification illustrating the encoding of information on one level.

In this example, the newly-created symbols ‘IC #’ mark the position in New where ‘i n f o r m a t i o n c o m p r e s s i o n’ was unified with the matching pattern in Old and thus deleted from New. And a copy of the symbol ‘IC’ is placed at the start of that sub-string in Old, together with a copy of the symbol ‘#’ placed as a ‘right boundary symbol’ or ‘termination marker’ at the end of the same pattern in Old.

---

5. In this context, the ‘sections’ are non-overlapping sub-strings within New. The length of each section is not pre-defined but is determined by the encoding process itself.

As indicated above, the code symbols might have been in place already in Old (as a result of some previous unification). In that case, copies of those symbols may be used directly in the encoding of New.

3.4.1 The format of coded patterns

The format for ‘coded’ patterns which are created as in the example above is this:

<LB><DS><DATA><RB>

where:

- ‘LB’ is a symbol like ‘(’ or ‘<’ marking the ‘left boundary’ of the pattern. Quite often, the ‘DS’ (described next) may serve as a left boundary marker so that the LB symbol may be omitted (as was the case in the example shown in Fig. 1).
- ‘DS’ is one or more ‘distinguishing’ or ‘discrimination’ symbols, like ‘IC’ in Fig. 1, which serves to differentiate the pattern from other patterns.
- ‘DATA’ is one or more ‘data’ symbols like the characters ‘i n f o r m a t i o n c o m p r e s s i o n’ in Fig. 1.
- ‘RB’ is a symbol like ‘)’, ‘>’, ‘#’ or ‘;’ which is a ‘right boundary’ or ‘termination symbol’ marking the end of the pattern.

In the remainder of this paper, symbols which function as LB or RB will be referred to as ‘boundary symbols’. The three classes which are not ‘DATA’ will be referred to collectively as ‘code’ symbols or ‘encoding’ symbols.

3.5 The formation of disjunctive groupings and hierarchical ‘grammars’

A ‘good’ alignment of two patterns may easily have one or more gaps (unmatched symbols) in the sequence of hits between the two patterns. For example, ‘t h a t b o y d a n c e s’ and ‘t h a t g i r l d a n c e s’ can be matched at the ends but not in the middle. With this example, the unified pattern may become something like ‘S t h a t N ; d a n c e s #’, together with the patterns ‘N 0 b o y ;’ and ‘N 1 g i r l ;’. The last two of the patterns just shown constitute a disjunctive class which is a sub-set of the class of ‘nouns’.

The three patterns just shown constitute the beginnings of a ‘grammar’ which is hierarchical in the sense that any one pattern may itself contain code symbols for other patterns, which may themselves contain codes for other patterns - and so on through arbitrarily many ‘levels’. There may be conjunctive groupings (‘chunks’), or disjunctive groupings (‘classes’), or both, at any level.

It is not appropriate here to describe or discuss in more detail how such structures may be built up. The main point here is that the ICMAUS framework appears to accommodate the formation of hierarchically-encoded structures which are functionally equivalent to grammars. Further information about how structures of this kind may be learned may be found in [13, 14, 15, 20, 21].

3.6 Parsing, encoding and decoding using an hierarchical grammar

The SP51 model is dedicated to the parsing of patterns in New in terms of pre-established grammars of the kind just considered (in Old) using the principles of ICMAUS. A detailed account of how the model may achieve parsing through arbitrarily many levels is presented in [10].

From a successful parsing of a pattern, a compact encoding of the pattern in terms of the grammar may be derived. How this can be done is described briefly in Section 4.0.3, below, and more fully in [10].

Curiously, the *decoding* of an encoded description of a pattern may be achieved by *exactly* the same ICMAUS mechanisms as were used to achieve encoding. How this may be achieved with the SP51 model is described and discussed in [10].

3.7 Identification of discontinuous patterns

One of the attractions of the ICMAUS framework is that it accommodates the matching of sub-sequences within patterns (in which the constituent symbols need not be contiguous) as well as sub-strings. For example, 'A X Y B Z C' may be aligned with 'L A M N O B P Q R S C T' yielding the shared sub-sequence 'A B C'.

If lossy compression is acceptable, as it often is, then the ICMAUS framework provides an effective means of discovering redundancy in the form of discontinous or fragmented patterns (see [15, Section 5.5]).

The way in which lossless compression may be achieved in cases like this would take us too far afield and will not be discussed here.

4 QUANTIFICATION OF COMPRESSION

Methods of calculating the information compression associated with any alignment are described in [1-4]. However, these methods are not suitable for the alignment problem as it has been framed here. This section describes the method of calculating the compression associated with any alignment which is used in SP51.

The method described here is applicable only in cases where an alignment can be interpreted as a parsing or analysis of New in terms of patterns in Old. The main change which will be required when the model is generalised to include the learning of new patterns is a method of calculating the information cost of new code symbols which are introduced as learning proceeds. This is discussed in Section 4.0.5, below.

In SP51, the compression score for any alignment, C, is calculated as:

$$C = B_n - B_e \qquad (1)$$

where B_n is the number of bits required to represent the New information in the alignment in 'raw' form without any encoding (except, possibly, for Huffman coding (see, for example, [5, 8]) or Shannon-Fano-Elias (SFE) coding (see [5]) at the level of single symbols), and B_e is the number of bits required for the encoding of New information in terms of Old information. How these values are calculated is described below.

4.0.1 *Information costs of symbols*

If a simple block code is used for symbols, then the 'minimum cost', M, for each symbol is

$$M = \lceil \log_2|S| \rceil \qquad (2)$$

bits where |S| is the number of symbol types in the alphabet of symbol types (S) used throughout New and Old.

In SP52 the value of M for each symbol type (and thus each individual symbol) is calculated by the SFE method (described in [5]), which assigns short codes to frequently occurring symbols and longer codes to rarer ones.[6]

Whichever way the value of M is calculated, the 'actual cost', A, of each symbol is

$$A = \lceil M \times c \rceil \qquad 1 < c < 1.5 \qquad (3)$$

where c is a factor which is normally only a little greater than 1.

The reason for the distinction between 'minimum cost' and 'actual cost' is described in [10] and will not be repeated here because it is not relevant to the subject in hand.

Notice that M and A are totally independent of the strings of characters which are used to represent the symbols in examples. The latter are chosen purely for their mnemonic value and to aid understanding and do not reflect the number of bits required for matching and discrimination of symbols.

4.0.2 *Calculation of* B_n

B_n is the number of bits required to express New in its 'raw' form, without any encoding except, possibly, SFE coding of individual symbols. B_n is calculated as:

$$B_n = \sum_{i=1}^{n} A_i \qquad (4)$$

where A_i is the 'actual cost' of the *i*th symbol within the sub-sequence from New whose symbols form hits with symbols in Old in the current alignment.

4.0.3 *Derivation of a code from a parsing and the calculation of* B_e

For any parsing or partial parsing, there is a corresponding code - which is a more or less compressed representation of the sentence (or part of a sentence) which has been parsed. The code for any parsing is the code symbols for the patterns from Old which appear in the alignment *less* any code symbols from Old which are matched within the parsing. For example, the phrase '`t h e b o y`' may be parsed using the patterns '`D 1 t h e ;`' and '`N 0 b o y ;`' which, in turn, may be aligned with the pattern '`NP D ; N ; #`'. With the first two

---

6. The SFE method produces codes which are slightly less compact than codes produced by the Huffman method. However, SFE codes provide a more accurate reflection of the relative frequencies of entities and this is exploited in the methods of calculating probabilities which are described later.

patterns, the code is ‘D 1 ; N 0 ;’. But with the inclusion of the third pattern, the code may be reduced to ‘NP 1 0 ;’. This is because ‘NP’ signals that the structure is a noun phrase and, for that reason, the sequence of symbols ‘D ; N ;’ is redundant and may be removed from the sequence ‘D 1 ; N 0 ;’. Fuller discussion may be found in [10].

Following the principles which have just been outlined, the number of bits required to encode a given alignment is calculated as:

$$B_e = \sum_{j=1}^{p} (E_j \cdot a_j) - \sum_{k=1}^{h} M_k \qquad (5)$$

where E_j is the number of bits required as the code for the jth pattern out of the p patterns in Old which appear in the alignment (calculated as described in Section 4.0.4, below), a_j is the number of times the jth pattern from Old appears in the given alignment and M_k is the minimum cost of *one* of the two (identical) symbols in the kth out of h hits in the alignment between two ‘code’ symbols in Old.

4.0.4 Calculation of E

The calculation of B_e for any alignment requires a value for the ‘encoding cost’, E, for each pattern from Old which appears in the alignment (all of which are assumed to have been ‘coded’). This is the sum of the bits in the ‘code’ symbols described in Section 3.4.1:

$$E = M_{LB} + \sum_{d=1}^{r} M_{DS} + M_{RB} \qquad (6)$$

where M_{LB} is the minimum cost of LB (if there is one), M_{DS} is the number of bits required for the dth of r ‘discrimination’ symbols, DS, and M_{RB} is the minimum cost of the RB symbol.

4.0.5 Taking account of the frequencies of occurrence of patterns

In accordance with the SFE principle, the number of bits which are used in the code for each pattern should reflect the frequency of occurrence of each pattern: frequently occurring patterns should receive short codes while rarer patterns should receive longer codes. To be more specific, this principle applies to the ‘discrimination’ symbols in each code (CS and SS), not to the code symbols which serve as boundary markers. More detail is given in Section 6.2.

When a match between two patterns is first observed, we have no means of knowing what its long-range frequency of occurrence is likely to be. Initially, we should give discrimination symbols to that pattern which are sufficiently distinctive to discriminate the pattern uniquely from all other patterns stored in Old. However, the frequency of occurrence of each pattern should be recorded as the Corpus grows so that the discrimination symbols which have been assigned to patterns can be periodically reviewed and re-assigned in accordance with the SFE principle.

This idea will be examined further when the SP52 model is generalised to include the learning of new patterns and the assignment of code symbols to each pattern (whilst retaining the ability to parse New patterns in terms ones already stored in Old).

4.0.6 *Taking account of the sizes of patterns*

As was noted in Section 3.2.3, IC depends not only on the frequencies with which patterns occur but also on their *sizes*: for a given code cost, the unification of big patterns will yield more compression than the unification of small patterns.

Any simple-minded search which attempts always to maximise the sizes of patterns being unified at the earliest opportunity will, on occasion, lead to sub-optimal 'garden path' errors in the matching of patterns in New to patterns in Old. Other things being equal, the system should attempt to maximise the *average* sizes of patterns over a reasonable span and thus avoid the simpler kinds of garden-path errors.

By seeking to maximise C over a reasonable span, the SP52 model tends automatically to maximise the average sizes of patterns which are unified.

5 LEARNING OF CAUSAL CONNECTIONS AS ICMAUS

How can any child, adult or artificial system learn the connection between a set of causal conditions and the corresponding effect? This section considers what may be required and how the ICMAUS framework might be applied.

In what follows, it will be assumed, contrary to fact, that we live in a one-dimensional world - in keeping with our previous adoption of one-dimensional patterns as the stuff from which alignments may be constructed. If ICMAUS concepts prove useful in this simplified form, we may feel encouraged, in the future, to generalise them to two or more dimensions.

Since some features of objects in the world have no intrinsic order, and since the one-dimensional ICMAUS framework assumes that all patterns are ordered, we shall make the additional simplifying assumption that any features which have no intrinsic order will be given an arbitrary order to ensure consistency from one pattern to another.

5.1 Causality and correlation

From the definition of causality which we have adopted, it is difficult to resist the conclusion that the process of learning a causal connection is a process of observing that whenever the set of conditions in C obtain, E always occurs at the same time or later. In addition, there must be the inductive prediction, about which so many philosophical words have been written, and about which no additional words will be offered in this paper, that past observations are a guide to the future.

5.1.1 *The correlation paradox*

Identifying a causal link by observing that, whenever C occurs, E always occurs sounds suspiciously like the fallacy of arguing from correlation to causation. Every statistician and experimental scientist knows that a correlation between two things is an unreliable basis for inferring a causal relation between them: it can easily happen that both things have a common cause. The behaviour of two radios which are tuned to the same station may be perfectly correlated without anyone supposing that the sounds coming from one radio are causing the same sounds to come from the other.

How can we resolve this seeming paradox? On the one hand, the concept of causation which we have adopted seems to be equivalent to a form of strong correlation between events, and on the other hand, we know that correlations between events are not a reliable basis for inferring causal connections!

5.1.2 Resolving the correlation paradox

There seem to be two ways of resolving this paradox which are not mutually exclusive.

The first consideration is the 'window' which we have on the relevant or potentially relevant information. If we have no knowledge of radios or transmitters and we can merely listen to two radios without any opportunity to experiment with them, then we might well suppose that a causal connection between them was at least a possibility, whilst acknowledging that there might be a third causal agency controlling them both. However, if our window onto the world of radios were expanded to include knowledge of transmitters and if we had opportunities to experiment with the radios and transmitters (thus expanding the relevant information available to us), then our understanding is likely to change in favour of the transmitter as the causal factor.

The second consideration is the 'strength' of relevant correlations. Given our expanded window onto the world of radios and transmitters, we would quickly discover that either of the two radios could be switched off without preventing the other one from working and without affecting the transmitter. By contrast, switching off the transmitter would always stop both radios and switching it on again would start them both again (assuming that both radios are switched on and tuned to the transmitter). Evidence like this would show that there is a *stronger* correlation between the behaviours of the transmitter and each radio than there is between the behaviours of the two radios. Since the strong correlations between the transmitter and each radio account for the weaker one between the two radios, it is appropriate to identify the transmitter as the causal factor (together with the provision of electricity and other 'background' factors which are the same in all cases).

The gist of the argument is that we can salvage the idea of causation as a form of correlation or co-occurrence of events if we have access to a sufficiently large body of relevant information and if we allow 'competition' between alternative causal explanations, selecting the one which is supported by the strongest correlation or correlations.

An implication of this view is that, for *any* putative causal connection, there is always the possibility that, with additional information, some other *stronger* correlation may be discovered which would may be adopted as the basis for a new 'correct' causal explanation.

5.2 Correlations and causal connections as patterns

If the foregoing argument is accepted, every causal connection may be understood as a kind of correlation or regularity in the world.

Correlations and regularities are forms of redundancy. So, in accordance with the working hypothesis described in Section 3.2.3, we may suppose that any correlation of the kind we have been discussing may be expressed in the form of *patterns*. Here are some examples.

Let us suppose, for the sake of clarity and simplicity, that the sounds coming from our two radios ('r1' and 'r2') are pure tones of fixed length of only two pitches, 'high pitch' (hp) and 'low pitch' (lp), each tone separated from the next by a silence (s). The correlation between the two radios may be expressed with three patterns like these:

P 1 r1 hp , r2 hp ;
P 2 r1 lp , r2 lp ;
P 3 r1 s , r2 s ;

These patterns express the fact that when the first radio is producing 'high pitch' the second one is too, and likewise for 'low pitch' and 'silence'. The correlations would be equally good with 'P1 r1 hp r2 lp', 'P2 r1 lp r2 hp' and 'P3 r1 s r2 s', or any other recurring combination.

When the transmitter (t) is brought into the picture (assuming that the radios are switched on and tuned to the transmitter), the patterns become:

P 4 t hp , r1 hp , r2 hp ;
P 5 t lp , r1 lp , r2 lp ;
P 6 t s , r1 s , r2 s ;

where, in the context of the transmitter, 'hp', 'lp' and 's' may be taken to represent input values of the sounds to be transmitted. In the light of the remarks in Section 4.0.6 about redundancy and the sizes of patterns, it is clear that these three patterns represent more redundancy than the three above. Since these three patterns include the information in the three above, they may be chosen as a better representation of regularities in the world of radios and transmitters than the first three patterns.

In the case of the 'on/off' behaviour of the radios and the transmitter (t), the regularities may be expressed with patterns like these:

P 7 t on , r sw-on on ;
P 8 t off , r sw-on off ;
P 9 t on , r sw-off off ;
P 10 t off , r sw-off off ;

where 'r' means any radio tuned to the transmitter, 'sw-on' and 'sw-off' means that the radio is switched on or off, and 'on' and 'off' have their obvious meanings.

5.2.1 Other redundancies

Readers may, with some justice, object that, while the representation of these regularities as patterns makes reasonable sense when the problem is simplified, it would not look nearly as good in more realistic cases where, for example, the sound waves were to be plotted with a resolution of, say, 1000 points between the highest peak and the lowest trough. In this case, it would require 1000 patterns to express the correlation between the two waves.

In a case like this, we would normally avoid using 1000 patterns to express the correlation between the two radios (and similar correlations between the transmitter and each radio) by taking advantage of the fact that there is an additional regularity in the data: whatever the value is at one location, the corresponding value at any other location is the same. Using conventional notations, we would express this in some mathematical form which would be much more compact than the use of 1000 patterns.

It appears that the ICMAUS framework as outlined in this article is capable of capturing this kind of redundancy in a manner which is comparable with the way it would be done with mathematical notation. How this can be done is discussed elsewhere [11, 12]. It would take us too far afield to discuss it here and it would in any case not be germane to understanding causality where things which are causally connected are, in general, not the same.

Notice that the correlations amongst the transmitter and the two radios would still be perfect even if some scrambling system consistently substituted some arbitrary value at one location for the corresponding value at any other location. In this case, the use of 1000 patterns to represent the correlation (each pattern showing a correspondence between values) would be unavoidable in any system, conventional or otherwise.

5.2.2 Asymmetrical regularities

In Section 2, it was noted that there is a fundamental asymmetry in the concept of causality which we have adopted: an effect may have two or more alternative causes but there is only one effect or set of effects for any cause.

At first sight, this asymmetry in causality seems incompatible with the idea of recording regularities in the form of patterns - because each pattern is merely a sequence of symbols without any asymmetry amongst those symbols.

The example, above, of the on/off behaviour of a transmitter and a radio gives an indication of how this seeming anomaly can be overcome: using patterns, asymmetrical relations can be represented by employing two or more patterns to represent the asymmetry. In the case of the transmitter and the radio, the four patterns, above, show that the on or off state of the transmitter has an effect on any radio which is switched on and tuned to that transmitter but switching the radio on or off has no effect on the transmitter.

Here is another example related to car engines:

```
F 1 ewng nsp ;
F 2 ewng nf ;
F 3 ewng na ;
F 4 ewng nsp na ;
```

where ‘F 1’ to ‘F 4’ are arbitrary distinguishing patterns, ‘ewng’ means “engine will not go”, ‘nsp’ means “no spark at spark plugs”, ‘nf’ means “no fuel reaching the engine”, and ‘na’ means “no air reaching the engine”.

These patterns represent four different associations between the engine of a car failing to go and possible explanations. Section 6 discusses how asymmetrical ‘probabilities’ may be derived from patterns like the ones just shown and, correspondingly, why we may feel entitled to regard the relationships represented in these patterns as causal. For the time being, we will take it on trust that patterns of the kind described above do indeed provide an effective way of representing causal relationships.

5.3 How causal patterns may be learned by ICMAUS

Given our simplifying assumption that we live in a one-dimensional world, it is not hard to see how patterns of the kind which were described in the last sub-section may be identified by the alignment of patterns representing 'raw' data.

We will assume that our learning system is already familiar with cars and car engines which are in working order. Given this prior knowledge, a single instance of a car containing an engine which had stopped working where there was also no spark at the spark plugs could be sufficient to identify an association between an engine not working and the absence of any spark at the spark plugs. This is because associations can be derived from clusters of non-hits as well as clusters of hits - as was the case where 'b o y' and 'g i r l' were picked out as discrete chunks from the alignment and unification of 't h a t b o y d a n c e s' and 't h a t g i r l d a n c e s'.

Of course, it might happen in that single case that the car in question also had a broken windscreen. In that case, the naÏf learning system, with no knowledge of physical or engineering principles, would conclude that the broken windscreen was part of the cluster of features associated with the engine not working. However, experience of some more cases of engines not working because of an absence of spark at the plugs would tend to exclude such things as broken windscreens and missing door handles because they would not be consistently associated with the engine failing to go.

In the discussion so far, we have assumed implicitly that features which are associated in a statistical sense are also contiguous, one with another, in the patterns which describe individual cars. In general, of course, this will not be true: most of the statistical associations in which we are interested are likely to be between features which do not form substrings within the patterns describing a car. In general, these associations will be 'discontinuous' in the sense described in Section 3.7. However, this does not invalidate the foregoing arguments. As was noted in Section 3.7, the redundancy represented by repeating but discontinous patterns can be detected and extracted in the ICMAUS framework, especially if lossy compression is acceptable, as it would be if our primary interest is in the recurrent associations and not the original patterns whence they came.

Fig. 2 shows an alignment of patterns describing cars in which failure of the engine is associated with absence of a spark at the plugs. From this alignment a single pattern may be derived by unification, 'F 1 ewng nsp ;', which records the association.

Fig. 2. An alignment of patterns describing cars which shows a regular association between failure of the engine and no spark at the plugs.

6 CAUSAL INFERENCES AS ICMAUS

Given that causal connections have been learned and recorded as coded patterns in Old, how may they be used in the ICMAUS framework for the making of inferences about causes and effects? This section considers this question in two stages, first examining how relatively simple 'one-step' inferences may be made and then more complicated kinds of inference where there may be two or more steps in the inference.

Unlike previous discussions of inference via ICMAUS [11, 13], the presentation here includes a quantitative analysis of 'probabilities' including a demonstration of how the asymmetry of causal connections may be quantified.

6.1 One-step inferences

Each of the four patterns, above, describing associations between causes and effects in the working of a car engine may be associated with a number representing the number of times that that association has been observed. In the context of ICMAUS, this 'frequency of occurrence' means the number of times that a pattern from New has been unified with the given pattern in Old.

Here are the four patterns, each one with the kind of value for frequency that one might obtain in practice:

```
F 1 ewng nsp ;       24
F 2 ewng nf ;        33
F 3 ewng na ;        15
F 4 ewng nsp na ;    1
```

Numbers like these may be used in a straightforward manner to calculate conditional probabilities of causes and effects as will now be described using these examples.

6.1.1 Inferences from effect to cause

Consider first of all, how, in the ICMAUS framework, one may make inferences from an effect to possible causes of that effect. In the following discussion, the terms 'cause' and 'effect' will be used as convenient shorthands but, of course, the corresponding concepts have no formal status in the ICMAUS framework.

To achieve inferences from an effect to possible causes of the effect, we need the four patterns above in the 'Old' part of the corpus *together with all other patterns which are relevant to causal relations in the workings of car engines*. In addition, we need a 'query' pattern in the 'New' part of the Corpus such as 'F ewng ;' which contains the symbol 'ewng' representing the effect in which we are interested.

With patterns like these, an ICMAUS search will find alternative 'good' alignments like the four shown in Fig. 3.

Fig. 3. Four alternative alignments between 'F ewng ;' (in New) with four patterns contained in Old.

The data symbols in each alignment which are *not* matched to any symbol in 'F ewng ;' constitute answers to the query "What data symbol or symbols are associated with the symbols 'F ewng ;'?", or "What are the possible causes of the engine of a car not going?"

Of course, the alternative answers include 'nsp' ('no spark'), 'nf' ('no fuel'), 'na' ('no air') and (relatively rarely) a combination of the first and the last of the faults just mentioned. Other answers amongst the alternatives would be "seized pistons", "broken crankshaft", "badly-set timing", "leaking valves" and so on.

The probability of each alternative is clearly the frequency of the pattern in which the given alternative appears divided by the total frequency of the alternative patterns which have been ‘selected’ by the query pattern (and which match all the symbols in the query pattern):

$$p_j = f_j / \sum_{i=1}^{n} f_i \qquad (7)$$

where p_j is the probability of the jth out of n alternative alignments, f_j is frequency of occurrence of the pattern which is ‘selected’ in the jth alternative, and f_i is the frequency of occurrence of the pattern which is selected in the ith alternative.

Let us assume that the four patterns from Old which appear in Fig. 3 are the complete set of patterns where each pattern matches all the symbols in the query pattern. The sum of the frequencies of these four patterns is 24 + 33 + 15 + 1 = 73. The corresponding probabilities for the four alternatives we have been considering are: ‘nsp’: 24/73 = 0.33, ‘nf’: 33/73 = 0.45, ‘na’: 15/73 = 0.21, and ‘nsp na’: 1/73 = 0.01.

6.1.2 Inferences from cause to effect

If we wish to discover the effect of a given cause, we may apply a query pattern like ‘F nsp ;’ which contains a symbol representing the cause in which we are interested.

In this case, there are two alternative alignments with patterns in Old, as shown in Fig. 4.

Fig. 4. Two alternative alignments between ‘F nsp ;’ (in New) with patterns contained in Old.

The first alignment shows that ‘ewng’ is associated with ‘nsp’. The second alignment also shows that ‘ewng’ is associated with ‘nsp’ but, in addition, it shows that ‘na’ is associated with ‘nsp’.

If we apply formula (7), above, we find that the probabilities of the first alternative is 24 / (24 + 1) = 0.96, and the probability of the second alternative is 1 / (24 + 1) = 0.04. However, we can ‘see’ that the symbol ‘ewng’ appears as an associate in both cases - so the probability of ‘ewng’ in the context of ‘nsp’ is 0.96 + 0.04 = 1.00. By contrast, the symbol ‘na’ appears in only one of the two alignments, so its probability in the context of ‘nsp’ remains at the low value of 0.04.

Similar alignments may be found and calculations made for the causal factors ‘nf’ and ‘na’ and the many other things which can prevent a car engine from going. In every case, there is an 100% chance that the engine will not go if the given causal factor is present.

6.1.3 Asymmetry in causality

In itself, each pattern showing an association between ‘cause’ and ‘effect’ is indifferent to questions of symmetry or asymmetry. But alignments (and unifications) amongst two or more patterns may yield probability values showing marked asymmetries: for any ‘cause’ (or combination of causes), there is a probability of 1.0 that the associated ‘effect’ will be inferred, but for any ‘effect’ there is a relatively low probability for each of the alternative possible ‘causes’. This asymmetry in probabilities derived from ICMAUS inferences accords with the observations made in Section 2 about the intrinsic asymmetry of the concept of causality which we have adopted.

6.2 Frequencies, probabilities and the sizes of codes

The foregoing discussion of probabilities assumes that each pattern in Old has a value attached to it representing its frequency of occurrence. The same arguments can be made if we suppose that each pattern has a value (P) attached to it representing the absolute probability of the pattern - in which case each value of f in formula (7) may be replaced by the corresponding value for P.

Where or how should these values for frequency or probability be stored? At first sight, the answer to this question seems to demand some storage attached to each pattern which is outside the pattern itself. However, a little reflection shows that the manner in which patterns are constructed allows information about absolute probability to be stored within the pattern without the need for any other mechanism or receptacle.

Within each pattern, there should be one or more symbols whose function is to identify the pattern, distinguishing it from all other patterns in the set of patterns in Old (Section 3.4). These symbols should be assigned to each pattern in accordance with the SFE method (described in [5]) so that common patterns are identified with fewer bits than rare patterns.

As previously noted, the SFE principle has been chosen in preferance to the Huffman principle because it provides a more accurate reflection of the frequencies of symbols and patterns.

Given the application of this principle, the relationship between the size of the discrimination symbols (in bits) and the probability of the pattern is:

$$s = \lceil \log_2 (1 / P) \rceil + 1 \qquad (8)$$

where s is the total size of the discrimination symbols for any pattern (in bits) and P is the absolute probability of the pattern. From this formula, or the equivalent formula,

$$P = 1 / (2^{s-1}), \qquad (9)$$

it is easy to obtain a value for P for any value of s.

6.3 Indirection in causal inferences

Causal inferences are not always as simple and direct as the examples discussed in Section 6.1. From time to time we make judgements about causation where there are two or more steps in the argument.

For example, we may observe that the battery of a car is flat. From this observation, we infer that there will not be any sparks at the spark plugs. And from this inference, we infer that the engine will not go. In a similar way, observation that the poles of the battery have been shorted may lead us to make the inference that the battery will be flat and from there infer that there will be no spark at the spark plugs and that the engine will not go.

Inferences like these may be made even if there is no pattern stored in Old showing an association between the initial cause and the final effect. We may have no direct knowledge that a flat battery is associated with the engine failing to go but we can piece this inference together from our knowledge that a flat battery will inhibit the production of sparks at the spark plugs and our knowledge that the absence of sparks means that the engine will not go.

It appears that the ICMAUS framework may accommodate 'chains' of inference like these in much the same way that it may accommodate grammars with arbitrarily many levels in the parsing and creation of sentences [10].

For example, knowledge which is relevant to the inferences just described may be represented with patterns like these:

```
F 1 ewng EF # ;
F 2 ewng nf ;
F 3 ewng na ;
F 4 ewng EF # na ;
...
EF 1 nsp damp distributor #
EF 2 nsp BF % #
EF 3 nsp coil faulty #
...
BF 1 flatb battery poles shorted %
BF 2 flatb no acid in battery %
BF 3 flatb battery too old %
...
```

where 'EF' is mnemonic for "electrical fault", 'BF' means "battery fault", 'flatb' is short for "flat battery", and the other symbols have their obvious meanings.

Amongst these patterns, the first one ('F 1') contains a 'reference' to patterns which start with 'EF' and end with '#'. Amongst those patterns, the second one ('EF 2') contains a 'reference' to patterns which start with 'BF' and end with '%'. Each of these references is a shorthand for a set of features which are alternatives in the given context.

If patterns like these are in Old, a 'query' may be applied to them (in New) in the form of a pattern like 'F battery poles shorted ;', which may be read as a request to find the (causal) implications of a shorting of the poles of a car battery. A search for alignments which will allow the encoding of all three symbols in the query pattern yields the alignment shown in Fig. 5 and a similar one in which pattern 'F 4' replaces pattern 'F 1'. This is confirmed by a run of SP52 using these data.

Fig. 5. An alignment of patterns modelling the causal implications of the shorting of the poles of a car battery.

Both these alignments show that shorting of the poles of a car battery is associated with loss of charge in the battery so that it becomes flat which, in turn, means that no sparks can be formed at the spark plugs - which means that the car will not go.

If the query pattern had been 'F ewng ;', it not hard to see that a relatively large number of alternative alignments could be formed which would allow the encoding of all three symbols in the query. These correspond to the relatively large number of things which will cause a car engine not to go.

6.4 Calculation of the probabilities of inferences derived from multiple alignments

How can we calculate the probabilities of inferences derived from alignments like those just considered? The method proposed here is a simple generalisation of the method described in Section 6.1 with a generalisation of the principles described in Section 6.2.

For any pattern in New, there are n alignments which contain hits for all the symbols in that pattern ($n \geq 0$). For each such alignment, an 'encoded' representation of New may be derived as described in Section 4.0.3. This encoded representation normally comprises one or more discrimination symbols followed by a termination symbol. For present purposes, it is the discrimination symbols which are required and, more specifically, the number of bits, s, which is the sum of the minimum costs of these symbols.

In accordance with equation (9), a value for P may be calculated from the value of s for each of the n alignments. Using these values as substitutes for corresponding values of f in formula (7) a probability value may be calculated for each alignment and thus for the inferences embodied in the alignment.

In general, the argument is that, if discrimination code symbols have been assigned to patterns in accordance with the SFE principle, then the *combination* of these discrimination symbols which appears in the coded representation of any alignment may be used to obtain an estimate of the relative probability of that alignment and the inferences which it embodies.

6.5 Completeness of the 'database'

The methods of calculating probabilities of causal inferences which have been presented assume that the 'database' of patterns stored in Old is, in some sense, complete. If we have reason to believe that there may be other relevant information in the 'real' world which is not stored in Old, then adjustments to the probabilities would be necessary. No attempt has yet been made to adapt the methods of calculation to accommodate incompleteness of the information stored in Old.

7 CONCLUSION

This paper has considered, in outline, how ICMAUS concepts may be used to model the learning of causal associations and the drawing of causal inferences. It seems that this framework of ideas holds sufficient promise in this area to deserve further attention and development.

ACKNOWLEDGEMENT

I am grateful to John Hornsby of the School of Electronic Engineering, University of Wales, Bangor, for advice on some of the equations presented in this paper.

REFERENCES

1 Allison, L and Wallace, C. S. (1994) The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *Journal of Molecular Evol*ution 39, 418-430.

2 Allison, L, Wallace, C. S. and Yee, C. N. (1992) Finite-state models in the alignment of macromolecules. *Journal of Molecular Evol*ution *35*, 77-89.

3 Allison, L. and Yee, C. N. (1990) Minimum Message Length encoding and the comparison of macromolecules. *Bulletin of Mathematical Biology 52 (3)*, 431-453.

4 Chan, S. C., Wong, A. K. C. and Chiu, D. K. Y. A (1992*)* Survey of Multiple Sequence Comparison Methods. *Bulletin of Mathematical Biology*, 54 *(*4), 563-598.

5 Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*, New York: John Wiley.

6 Hospers, J. (1967). Cause, determinism and freedom. In *An Introduction to Philosophical Analysis*, Second Edition, Englewood Cliffs, N.J.: Prentice-Hall, Chapter 5, pp. 279-348.

7 Solomonoff, R. J. (1964). A formal theory of inductive inference. Parts I and II. *Information and Control 7*, 1-22 and 224-254.

8 Storer, J. A. *(*1988) *Data Compression: Methods and Theory*. Rockville, Maryland: Computer Science Press.

9 Wallace, C. S. and Boulton, D. M (1968). An information measure for classification. *Computing Journal 11(2)*, 185-195

10 Wolff, J. G. (submitted). Parsing as information compression by multiple alignment, unification and search.

11 Wolff, J. G. (submitted). Logic as information compression by multiple alignment, unification and search.

12 Wolff, J. G. (submitted). Multiple alignment and computing.

13 Wolff, J. G. (1996). Learning and reasoning and information compression by multiple alignment, unification and search. In A. Gammerman (Ed.), *Computational Learning and Probabilistic Reasoning*, Chichester: John Wiley, pp. 67-85.

14 Wolff J. G (1995). Computing as compression: an overview of the SP theory and system. *New Generation Computing*, 13, 187-214.

15 Wolff, J. G. (1995). Computing as compression: SP20. To appear in *New Generation Computing*, 13, 215-241

16 Wolff J G (1994). A scaleable technique for best-match retrieval of sequential information using metrics-guided search*, Journal of Information Studies*, 20(1), 16-28.

17 Wolff, J. G.(1994). Computing and information compression: a reply. AI Communications, 7(3/4), 203-219.

18 Wolff, J. G. (1993). Computing, cognition and information compression. *AI*

Communications, 6(2), 107-127.

19 Wolff, J. G.(1991). *Towards a Theory of Cognition and Computing*. Chichester: Ellis Horwood.

20 Wolff, J. G.(1988). Learning syntax and meanings through optimization and distributional analysis. In *Categories and Processes in Language Acquisition* (Eds Levy, Y, Schlesinger, I M, and Braine, M D S) pp. 179-215, Hillsdale, N.J.: Lawrence Erlbaum. (Reproduced in Ref. 19, Chapter 2).

21 Wolff, J.G., (1982). Language acquisition, data compression and generalisation, *Language & Communication*. *2*, pp. 57-89. (Reproduced in Ref. 19, Chapter 3)

Fig. 1.

Alignment:

New

... t h a t i n f o r m a t i o n c o m p r e s s i o n i s ...

... s o m e i n f o r m a t i o n c o m p r e s s i o n w a s ...

Old

Unification:

... t h a t IC # i s ...

... s o m e IC i n f o r m a t i o n c o m p r e s s i o n # w a s ...

Fig. 2

```
car   8  ...  ewng  ...  nsp  ...  #
 |             |          |        |
car  22  ...  ewng  ...  nsp  ...  #
 |             |          |        |
car  50  ...  ewng  ...  nsp  ...  #
 |             |          |        |
car  88  ...  ewng  ...  nsp  ...  #
 |             |          |        |
car  95  ...  ewng  ...  nsp  ...  #
```

Fig. 3.

F ewng ;
F 1 ewng nsp ;

F ewng ;
F 2 ewng nf ;

F ewng ;
F 3 ewng na ;

F ewng ;
F 4 ewng nsp na ;

Fig. 4.

F nsp ;

F 1 ewng nsp ;

F nsp ;

F 4 ewng nsp na ;

Fig. 5.

```
F                                              battery poles shorted      ;
|                                                 |       |       |       |
F 1 ewng EF                                       |       |       |     # ;
           |                                      |       |       |     |
          EF 2 nsp BF                             |       |       |   % #
                     |                            |       |       |   |
                    BF 1 flatb battery poles shorted %
```