Information compression by multiple alignment, unification and search as a framework for medical diagnosis

J Gerard Wolff

CognitionResearch.org.uk

Abstract

This paper describes a novel approach to medical diagnosis based on the SP theory of computing and cognition. The main attractions of this approach are: a format for representing diseases that is simple and intuitive; an ability to cope with errors and uncertainties in diagnostic information; the simplicity of storing statistical information as frequencies of occurrence of diseases; a method for evaluating alternative diagnostic hypotheses that yields true probabilities; and a framework that should facilitate unsupervised learning of medical knowledge and the integration of medical diagnosis with other AI applications.

Key words:
information compression, multiple alignment, medical diagnosis,

1 Introduction

The problem of providing computational support for medical diagnosis has been approached from many directions including logical reasoning, fuzzy logic, set theory, rough set theory, if-then rules, Bayesian networks, artificial neural networks, case-based reasoning, support vector machines, perceptrons, possibility theory, and others.

This paper describes a novel approach to diagnosis based on the SP theory of computing and cognition. The main attractions of this approach are:

- A format for representing diseases that is simple and intuitive.

Email address: jgw@cognitionresearch.org.uk (J Gerard Wolff).
URL: www.cognitionresearch.org.uk/sp.htm (J Gerard Wolff).

- An ability to cope with errors and uncertainties in diagnostic information.
- The simplicity of storing statistical information as frequencies rather than conditional probabilities.
- A method for evaluating alternative diagnostic hypotheses that yields true probabilities.
- A framework that should facilitate unsupervised learning of medical knowledge and the integration of medical diagnosis with other AI applications.

It should be emphasised that the primary purpose of this paper is conceptual: to describe an approach to medical diagnosis that is significantly different from the main alternatives and with potential advantages compared with those alternatives. Although a prototype of the proposed new system exists, it is not yet a shrink-wrapped tool that is ready for immediate application.

Key elements of the SP theory are first described, just sufficient for present purposes. Section 3 describes how the theory may be applied to medical diagnosis, viewed as a process of pattern recognition. This section also discusses how the SP system relates to several aspects of the diagnostic process, including causal reasoning and the process of acquiring the knowledge that is needed for accurate diagnosis. Section 4 compares this new approach to medical diagnosis with some of the alternatives. The paper concludes with an outline of what still needs to be done in this programme of research and with a review of the main points that have been made.

2 The SP theory

The SP theory grew out of a long tradition in psychology that many aspects of brain function may be understood as information compression (see, for example, [1–4]). It is based on principles of *minimum length encoding* [1] pioneered by Solomonoff [5], Wallace and Boulton [6], Rissanen [7] and others (see also [8]). An overview of the theory is presented in [9] and more detail may be found in other papers cited there (see also [10]).

The SP theory has been developed as an abstract model of *any* system for processing information, either natural or artificial. In broad terms, the system receives 'New' information from its environment and transfers it to a repository of 'Old' information. At the same time, it tries to compress the information as much as possible by finding patterns that match each other and merging or 'unifying' patterns that are the same. [2] An important part of this process is the building of 'multiple

[1] An umbrella term for 'minimum message length encoding' and 'minimum description length encoding'.

[2] The term 'unification' in the SP theory means a simple merging of two or more identical

alignments' as described below.

To date, the main areas in which the SP framework has been applied are probabilistic reasoning, pattern recognition and information retrieval [11], parsing and production of natural language [12], modelling concepts in logic and mathematics [13], and unsupervised learning [14,15].

2.1 Representation of knowledge

In the SP system, *all* kinds of knowledge are stored as arrays of atomic *symbols* in one or two dimensions called *patterns*. In work to date, the main focus has been on one-dimensional patterns (i.e., sequences of symbols) but it is envisaged that, at some stage, the concepts will be generalised to patterns in two dimensions.

For present purposes, we may define patterns and symbols as follows:

- A *pattern* is a sequence of symbols bounded by end-of-pattern characters such as '(' and ')', not shown in the examples in this paper.
- A *symbol* is a string of non-space characters bounded by white space (space characters, line-feed characters and the like).
- Any symbol can be matched with any other symbol and, for any one pair of symbols, the two symbols are either 'the same' or 'different'. No other result is permitted.
- Symbols have no intrinsic meaning such as 'add' for the symbol '+' in arithmetic or 'multiply' for the symbol '×'. Any meaning attaching to an SP symbol takes the form of one or more other symbols with which it is associated in a given set of patterns.
- Each pattern has an associated integer value representing the absolute or relative frequency of occurrence of that pattern in some domain.

Despite the extraordinary simplicity of this format for representing knowledge, the way in which SP patterns are processed within the system means that they can model a wide variety of established representational schemes, including context-free and context-sensitive grammars, class-inclusion hierarchies, part-whole hierarchies, discrimination networks and trees, if-then rules, and others.

---

patterns to make one. This meaning is different from but related to the meaning of the term in logic.

2.2 Processing knowledge

A key part of the process of matching patterns is the building of 'multiple alignments', illustrated below. As in bioinformatics (whence the concept has been borrowed), a multiple alignment in the SP framework comprises two or more sequences of symbols arranged so that matching symbols are aligned. By contrast with the bioinformatics concept, one or more of the patterns has the status 'New' and all the rest are 'Old'. A 'good' alignment is one that allows the New pattern or patterns to be encoded economically in terms of the Old patterns, as described in [9].

In bioinformatics, it is generally understood that the search space of alternative possible alignments between two or more sequences is astronomically large and cannot be searched exhaustively. All practical methods for finding 'good' alignments amongst two or more sequences use heuristic methods such as 'hill climbing (or 'descent'), 'beam search', 'genetic algorithms' or the like. With methods like these, one can find good approximate solutions in a reasonable time but one can never be sure of finding the best possible solution unless the sequences are very short and very few.

Finding good multiple alignments in the SP system is no different. The process has been realised in the SP62 computer model (an improved version of SP61 described in [12] [3]) which uses a form of hill climbing to find good multiple alignments.

The process of building multiple alignments in the SP system provides a unified model for a variety of computational effects including fuzzy pattern recognition, best-match information retrieval, probabilistic and exact styles of reasoning, unsupervised learning, planning, problem solving and others, as described in [9].

The SP framework is Turing-equivalent in the sense that it can model a universal Turing machine [16] but it has much more to say about the nature of 'intelligence' than the Turing model of computing (or equivalent models such as lamda calculus [17] or the Post canonical system [18]).

2.2.1 Computational complexity

The time complexity of the SP62 model in a serial processing environment is approximately O($log_2 n \times nm$), where n is the size of the New pattern or patterns (in bits) and m is the total size of the patterns in Old (in bits). In a parallel processing environment, the time complexity may approach O($\log_2 n \times n$), depending on how

[3] The main difference between SP62 and SP61 is that each multiple alignment created by SP62 contains one *or more* New patterns, whereas each multiple alignment created by SP61 contains *one* New pattern and only one such pattern.

well the parallel processing is applied. The space complexity in serial or parallel environments is O(m). Further details may be found in [12].

In medical diagnosis, it seems reasonable to suppose that there will normally be a fairly small maximum for the number of signs and symptoms (abbreviated hereinafter as 'symptoms') exhibited by any one patient. Correspondingly, there should be a maximum size for the size of the set of New patterns (that are used to represent the patient's symptoms). If we take this to be a constant value for n, then in a serial processing environment the time complexity is approximately O(m) and in a parallel processing environment it may approach O(1).

3 Application of the SP system to medical diagnosis

To a large extent, medical diagnosis may be viewed as a problem of (fuzzy) pattern recognition: finding the best fit between a given set of symptoms for an individual patient and the symptoms associated with one or more diseases. However, causal reasoning also has a part to play when, for example, it is understood that a given disease is caused by a bacterial or virus infection.

This section presents an example showing how the SP system may be applied to medical diagnosis, viewed as a process of pattern recognition. The system may also support causal reasoning about medical problems and this is discussed briefly in Section 3.9, below.

Within this main section, Subsection 3.10 describes how the SP system may facilitate the acquisition of knowledge that is required for medical diagnosis.

3.1 Describing diseases using SP patterns

In the SP scheme, knowledge about diseases may be stored as patterns in a repository of Old information and the symptoms for an individual patient may be represented as a set of one or more New patterns.

A pattern in the store of Old information may represent one disease and its associated symptoms, or a combination of diseases (see Section 3.7.2, below), or it may represent a cluster of symptoms that tend to occur together in two or more different diseases (see Section 3.5, below). In addition, Old may include patterns that play supporting rôles (see Section 3.6, below).

The frequency value of each pattern may be used to represent the absolute or relative frequency with which a given disease or cluster of symptoms is found in a

given population. These figures may be derived from population surveys or they may be estimated by medical experts.

By way of illustration, Figure 1 shows three examples of such patterns, one describing the symptoms of chicken pox, another describing the symptoms of smallpox and the third describing the cluster of symptoms which is described as 'fever'. The number in brackets after each pattern is a very rough estimate of the relative frequencies of occurrence of the corresponding disease or condition. [4]

Each of these example patterns begins and ends with a pair of symbols '<disease> ... </disease>' which indicate that the pattern describes a disease or a cluster of disease symptoms. Within each pattern, there are similar pairs of symbols, each one marking the beginning and end of a 'field' which describes some aspect of the disease or cluster. For example, '<dname> Chicken Pox </dname>' provides the name of the chicken pox disease, '<skin> rash </skin>' describes one of its symptoms, '<causative_agent> chicken pox virus </causative_agent>' describes what causes the disease, and '<treatment> chicken pox treatment </treatment>' is a remarkably unhelpful description of how to treat the disease which would, of course, be much more detailed in a fully developed knowledge base.

Within the pattern for chicken pox, the field '<R2> fever </R2>' indicates that 'fever' is one of the symptoms of the disease. However, by contrast with other fields like those just mentioned, the symbol 'fever' is, in effect, a reference or pointer to a cluster of symptoms such as rapid breathing, flushed face and high temperature described in the third pattern in the same figure. In a similar way, '<R1> flu_symptoms </R1>' in the pattern for smallpox is a reference or pointer to another pattern, not shown in the figure, that describes a cluster of symptoms associated with influenza and flu-like diseases. The way in which pointers like these are dereferenced in the SP system will be seen in the next section.

Readers who are familiar with XML [19] will notice that pairs of symbols like '<disease> ... </disease>' or '<dname> ... </dname>' are rather like the start and end tags used to mark the elements of an XML document. However, by contrast with XML and related languages such as HTML, symbols of that kind have no formal status in the SP system and the styles of symbols are not defined within the system. Any convenient style may be used such as 'disease ... #disease' or 'disease ... %disease' and in some applications it is not necessary to provide any distinctive markers for the beginnings and ends of patterns or fields. The concept of 'field' has no formal status in the SP system.

[4] The figure for smallpox is clearly too high in the world today but it will serve for the purpose of illustration.

```
<disease> chicken_pox :
     <dname> Chicken Pox </dname>
     <R2> fever </R2>
     <appetite> normal </appetite>
     <chest> normal </chest>
     <chills> no </chills>
     <cough> no </cough>
     <diarrhoea> no </diarrhoea>
     <fatigue> no </fatigue>
     <lymph_nodes> normal </lymph_nodes>
     <malaise> yes </malaise>
     <muscles> normal </muscles>
     <nose> normal </nose>
     <skin> rash </skin>
     <throat> normal </throat>
     <weight_change> no </weight_change>
     <causative_agent> chicken pox virus </causative_agent>
     <treatment> chicken pox treatment </treatment>
</disease> (2500)

<disease> smpx :
     <dname> Smallpox </dname>
     <R1> flu_symptoms </R1>
     <appetite> normal </appetite>
     <chest> normal </chest>
     <diarrhoea> no </diarrhoea>
     <fatigue> no </fatigue>
     <lymph_nodes> normal </lymph_nodes>
     <malaise> no </malaise>
     <skin> rash with blisters </skin>
     <weight_change> no </weight_change>
     <causative_agent> smallpox virus </causative_agent>
     <treatment> smallpox treatment </treatment>
</disease> (5)

<disease> fever
     <breathing> rapid </breathing>
     <face> flushed </face>
     <temperature> <t1> </t1> </temperature>
</disease> (15000)
```

Fig. 1. Three SP patterns, one describing the symptoms of chicken pox, another describing the symptoms of malaria and the third describing the symptoms of fever.

3.2 Multiple alignment and medical diagnosis

The process of diagnosis may be modelled by the building of one or more multiple alignments. Figure 3 shows the best alignment created by SP62 with a set of New patterns shown in Figure 2 that describe ‘John Smith’ and his symptoms and a set of Old patterns like those shown in Figure 1 that represent diseases or aspects of diseases. By convention, the New pattern or patterns in any alignment is always shown in column 0 with the Old patterns in the remaining columns, one pattern

per column. Apart from this constraint, the order of the patterns in the alignment is entirely arbitrary.

```
<patient> John Smith </patient>
<face> flushed </face>
<appetite> poor </appetite>
<breathing> rapid </breathing>
<muscles> aching </muscles>
<chills> yes </chills>
<fatigue> yes </fatigue>
<lymph_nodes> normal </lymph_nodes>
<malaise> no </malaise>
<nose> runny </nose>
<temperature> 38-39 </temperature>
<throat> sore </throat>
```

Fig. 2. The set of New patterns supplied to SP62 for the example discussed in the text.

An alignment like this may be interpreted as the result of a process of recognition. In this case, the symptoms that have been recognised are those of influenza, as shown in column 2. The following subsections discuss aspects of the alignment and of this interpretation.

3.3 A 'framework' pattern

In an application like this, it is convenient but not essential to include amongst the Old patterns a 'framework' pattern like the one shown in column 1. This is a generalised pattern for diseases of all kinds that lists the main categories associated with diseases such as '<dname> </dname>' (the name of the disease), '<breathing> </breathing>' (the state of the patient's breathing) and '<temperature> </temperature>' (the patient's temperature), but it does not specify specific values for any category.

This framework pattern serves as an anchor point for symbols in other patterns and facilitates the formation of multiple alignments in accordance with the rules described in [9] and earlier publications.

3.4 The ordering of descriptors

In an application like medical diagnosis, it is not obvious that there is any intrinsic order to the symptoms of a disease or associated descriptors such as the name of the patient. In describing a patient's symptoms, it should make no difference whether 'high temperature' is mentioned before 'runny nose' or the other way round.

In the SP framework, each pattern that describes a disease or a cluster of symptoms necessarily imposes an order in which categories of descriptors are specified.

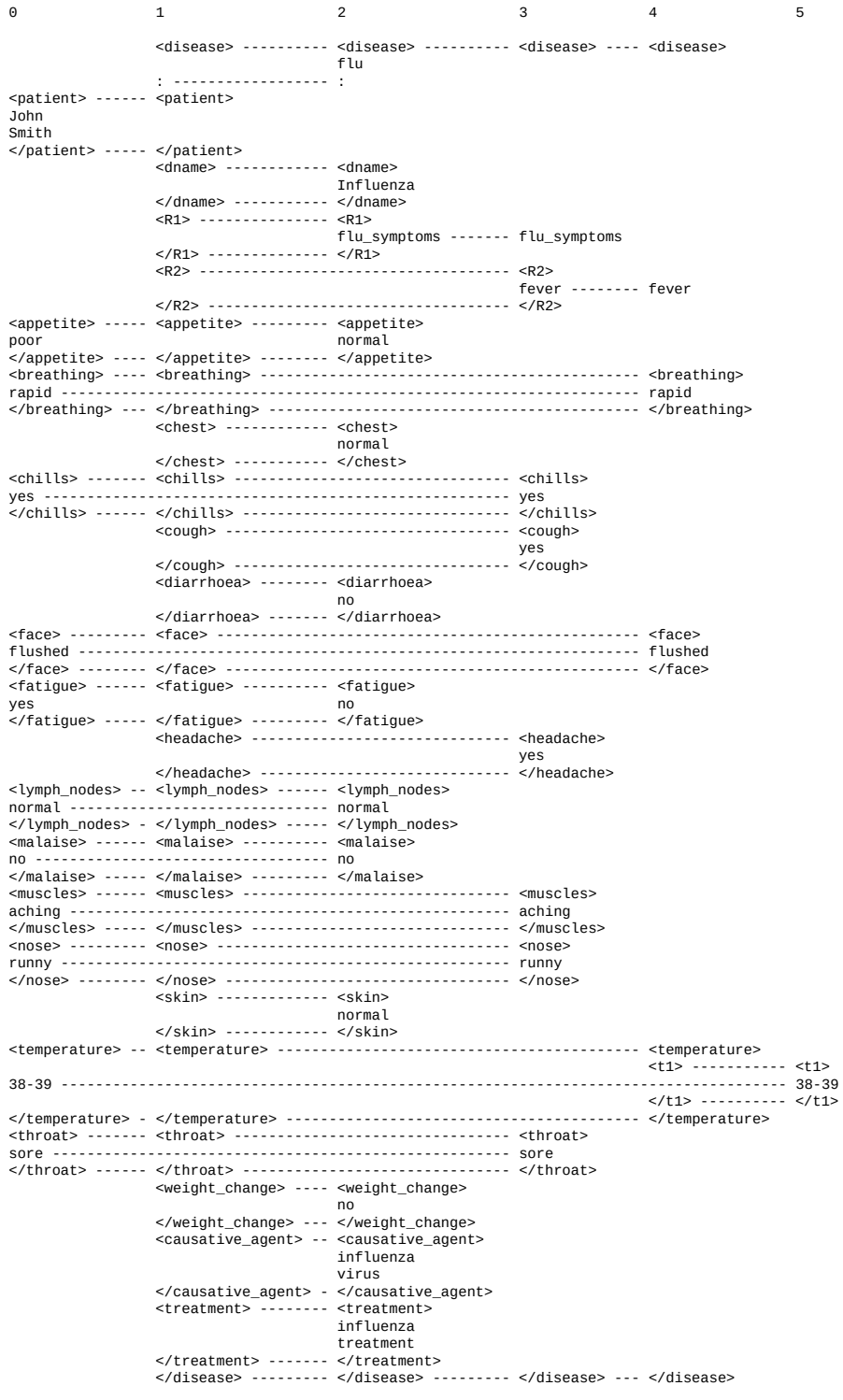

Fig. 3. The best alignment found by SP62 with the set of patterns from Figure 2 in New (describing the symptoms of the patient ‘John Smith’) and a set of patterns in Old describing a range of different diseases and named clusters of symptoms, together with the ‘framework’ pattern shown in column 1.

However, users of the system may specify the patient's symptoms in any order that is convenient. This is because symptoms are described using a set of New patterns and there is no intrinsic order amongst the New patterns supplied to the system. In our example, New patterns were supplied to the SP62 model in the order shown in Figure 2 but in the alignment shown in Figure 3 they appear in a completely different order.

Notice that this freedom in the ordering of descriptors only applies to whole patterns. When two or more symbols in one pattern are matched to two or more symbols in another, the order of the symbols in one pattern must be the same as the order of the matching symbols in the other pattern.

3.5 Dereferencing of pointers

As already noted, a symbol like 'fever' or 'flu_symptoms' in one pattern may serve as a reference or pointer to another pattern that describes a cluster of symptoms that may be found in two or more different diseases.

In Figure 3, we can see how such pointers are 'de-referenced' in the SP system. The symbol 'flu_symptoms' in column 2 is matched to the same symbol in column 3 where flu-like symptoms are listed. Likewise, the symbol 'fever' in column 3 is matched to the same symbol in column 4 where the symptoms of fever or listed. Fever is itself part of the cluster of flu-like symptoms.

The provision of named clusters like these saves the need to specify the corresponding symptoms redundantly in each of the diseases where such clusters appear.

3.6 Uncertainties in diagnosis

Diagnosis is not an exact process:

- Most diseases are 'family resemblance' or 'polythetic' concepts because the majority of symptoms associated with any given disease are neither necessary nor sufficient for the diagnosis of the disease: they are 'characteristic' of the disease in the sense that any one such symptom need not be present in every case and any of them may be associated with other diseases.
- There may be and frequently are errors in the observation or recording of symptoms.

SP62 can accommodate these kinds of uncertainty in diagnosis in two distinct ways:

- Because it looks for a global best match amongst patterns, it does not depend

on the presence or absence of any particular symptom. Notice how SP62 has succeeded in constructing the alignment shown in Figure 3 despite there being no match for 'poor' in the New pattern '<appetite> poor </appetite>' and 'yes' in the New pattern '<fatigue> yes </fatigue>' and no match for many of the symbols in the Old patterns.

Although the system does not depend on the presence or absence of any one symptom, particular symptoms can have a major impact on diagnosis, as described in Section 3.7.1, below.

- Within the SP framework, it is not necessary for every symptom of a disease to be recorded as a specific value. For example, in column 4 of the alignment in Figure 3, the pair of symbols '<t1> </t1>' represents a set of alternative values for the temperature associated with fever. In this case, there are just two values, represented in Old by the patterns '<t1> 38-39 </t1>' (high temperature) and '<t1> 40+ </t1>' (very high temperature). The first of these patterns is shown in column 5 of the alignment, matched to the temperature of the patient shown in column 0.

  Being able to specify symptoms as sets of alternative values allows the system to accommodate the kind of variability which is so prominent in many diseases.

3.7 Weighing alternative hypotheses and the calculation of probabilities

In medical diagnosis, it is quite usual for the physician to consider alternative hypotheses about what disease or diseases the patient may be suffering from. The SP framework provides a model for this process in the way the system builds alternative alignments for any given pattern in New. Alignments—and the corresponding diagnoses—may be evaluated as follows.

For each alignment, a 'compression difference' (CD) is calculated as $B_n - B_e$, where B_n is the total size (in bits) of those symbols within the New pattern that have been matched to Old symbols within the alignment, and B_e is the size (in bits) of the code for those New symbols, derived from the given alignment as explained in [9,12]. The details of how these values are calculated are explained in [12]. They are derived in part from the frequency values for patterns mentioned at the beginning of Section 3.2.

B_e may be translated into a probability:

$$p_{abs} = 2^{-B_e}. \tag{1}$$

For each alignment in a set of alternative alignments, $A_1 ... A_R$, that encode the same symbols from New, a relative probability may be calculated as:

$$p_{rel_i} = p_{abs_i} / p_{sum}, \tag{2}$$

where

$$p_{sum} = \sum_{i=1}^{i=R} p_{abs_i} \tag{3}$$

A fuller account of the way probabilities are calculated may be found in [11].

Given that the New patterns represent the symptoms of one patient at a particular time and given that each pattern in Old describes a single disease or a single cluster of symptoms that may form part of the description of one or more diseases, then each alignment formed by SP62 represents a hypothesis about any *one* disease that the patient may have.

Where alternative alignments encode different subsets of the symbols in New, it is possible that the patient may be suffering from two or more diseases at the same time. This possibility is discussed in Section 3.7.2, below. However, where two or more of the best alignments encode exactly the same symbols from New, then they represent *alternative* diagnostic hypotheses and they may be compared using values of p_{rel}.

When SP62 formed the alignment shown in Figure 3, it also formed a similar alignment, matching exactly the same symbols in New, in which column 2 contained a pattern representing the symptoms of smallpox, instead of the pattern for influenza. The p_{rel} values calculated in this case were 0.99950 for influenza and 0.00049 for smallpox, reflecting the prevalence of those two diseases in the world today. [5]

3.7.1 'Explaining away'

The symptoms of influenza and smallpox are quite similar, except for the very distinctive rash and blisters that occur in smallpox. The example shown in Figure 3 is silent about whether John Smith had a rash and blisters or not. If a rash and blisters had been seen to be absent, this would have been represented as '<skin> normal </skin>'. Given this lack of information about those symptoms, the patient is very much more likely to have influenza than smallpox, as indicated by the calculated probabilities.

If '<skin> rash with blisters </skin>' is added to the symptoms recorded in New, and if SP62 is run again with the augmented set of symptoms, the best alignment found by the system is similar to that shown in Figure 3 but with the pattern for smallpox (the second pattern in Figure 1) instead of the pattern for influenza in column 2 and with a match shown between '<skin> rash with blisters </skin>' in

[5] There are, of course, other factors that may be relevant—such as the possibility that someone might release the smallpox virus deliberately—but in this example knowledge of such other factors has been excluded.

the set of New patterns and the same symbols in the pattern that describes smallpox. However, in this case *there is no other alignment that matches the same symbols in New*. Consequently, the value of p_{rel} for the best alignment is 1.0. In short, the addition of one distinctive symptom to the list of symptoms has a dramatic effect on the relative probabilities calculated by the system. Instead of a vanishingly small probability for smallpox (0.00049), the system now assigns it a probability of 1.0, in accordance with our intuitions.

From this result, we may conclude that the patient certainly has smallpox and that his aching muscles and runny nose are due to smallpox, not influenza. This is the phenomenon of 'explaining away': "If A implies B, C implies B, and B is true, then finding that C is true makes A *less* credible. In other words, finding a second explanation for an item of data makes the first explanation less credible." ([20, p. 7], with the emphasis as in the original).

3.7.2 A patient may suffer from two or more diseases at the same time

As noted above, it is possible for a patient to suffer from two or more diseases at the same time. Given that the Old patterns in the system describe single diseases, then the system would create two or more 'good' alignments, each one corresponding to one of the diseases that the patient is suffering from.

If we want the system to calculate probabilities for combinations of diseases, then the repository of Old patterns must contain patterns that represent combinations of that kind. Each such pattern may be constructed economically using references to the component diseases, in much the same way that clusters of symptoms may be referenced, as described in Section 3.5.

As with single diseases, frequency values for a combinations of diseases may be obtained from population surveys or by the judgement of medical experts. In the absence of any direct evidence of a statistical association between two or more diseases, it seems reasonable to assume that they are statistically independent. In such cases, frequency values may be derived straightforwardly via normalised values for the frequencies of occurrence of individual diseases. Whether the frequency values for combinations of diseases are measured, estimated or derived, they can be used for the calculation of CD values and probabilities in exactly the same way as for single diseases.

Of course, there are so many possible combinations of diseases that it would be impossible to store information about them all. A more practical option may be to store information in Old about individual diseases and combinations of diseases that are known to have a statistical association with each other. One may assume that all other combinations of diseases are statistically independent.

3.8 Inferences and the diagnostic cycle

In a multiple alignment like the one shown in Figure 3, any symbol within an Old pattern that is *not* matched to a symbol in New represents an inference that may be drawn from the alignment. In this example, we may infer from the alignment *inter alia* that the patient is likely to have a cough and a headache and that the standard treatment for influenza is required. Probabilities of these inferences can be calculated as described in [11].

If a 'good' alignment makes a prediction about some marker that may be found in the patient's blood or something that may be observed in an X-ray, this may be interpreted as a suggestion to the physician that he or she should order an appropriate blood test or X-ray. If tests of that kind or other kinds of investigation are instigated as a result of the inferences drawn from preliminary alignments, the results of those investigations, together with the original symptoms, may be fed back into the system as New information. The system may then be run again and the alignments that are created may suggest a final diagnosis or the need for further investigation—and so on.

3.9 Causal reasoning

Apart from the kinds of inference just described, medical diagnosis often seems to involve a 'deeper' kind of reasoning about the causes of symptoms and diseases, using knowledge of bacteria, viruses, anatomy, physiology and so on.

The SP framework supports a variety of styles of reasoning, including probabilistic 'deductive' reasoning, abductive reasoning, nonmonotonic reasoning and (as we saw in Section 3.7.1) 'explaining away' (see [11]). So there are reasons to believe that, within the SP framework, it may be possible to extend the pattern recognition analysis described above to incorporate causal styles of reasoning.

Recent investigation has confirmed this expectation. The input-output relations of each subsystem within a larger system can be modelled in the SP framework as a set of patterns, and causal connections can be established by matching outputs to inputs. As with the analysis described above, a 'framework' pattern is also needed to ensure that alignments can be formed in an appropriate manner. These potential applications of the system need further exploration and development.

3.10 Acquisition of knowledge

Broadly speaking, the knowledge that is required in any artificial system for medical diagnosis can be obtained 'manually' from experts or written sources, or it may be obtained by the automatic or semi-automatic abstraction of knowledge from raw medical data, or some combination of the two. The SP system has potential to facilitate any or all of these processes.

3.10.1 Elicitation of expert knowledge

It should be apparent from the example described above that the SP system provides a means of representing medical knowledge in a form that is simple and intuitive. The simplicity of representing all knowledge as patterns is, perhaps, less important than the fact that this system allows computer-based knowledge to be expressed in a form that apparently reflects the natural structure of the original concepts.

This feature of the system should facilitate traditional kinds of knowledge elicitation from experts or written sources. Medical experts should have little difficulty in expressing their knowledge directly in the form of SP patterns. Given that such experts are often busy and their time is, in any case, expensive, there are advantages if at least some of the process of building computer-based knowledge bases can be undertaken by knowledge engineers without specialised medical training. It should be possible for such people to derive a good deal of the necessary knowledge from medical text books and other written sources.

3.10.2 Unsupervised learning

At its most abstract level (Section 2), the SP model is conceived as a system that learns by transferring New information to its repository of Old information and compressing it at the same time. This abstract conception has now been realised more concretely in the form of the SP70 computer model [14,15] that is capable of learning simple grammars from raw data. However, further development of the model is needed to realise its full potential.

The current model has two stages:

(1) From partial alignments between patterns, the model creates new patterns that are added to the repository of Old patterns. For example, if 'A B C P Q R S D E F G' is aligned in the obvious way with 'A B C X Y Z D E F G', the system isolates the matched subsequences 'A B C' and 'D E F G' as discrete elements and adds system-generated 'code' symbols to each one to create patterns like '<%1> A B C </%1>' and '<%2> D E F G </%2>'. In a similar way, the system isolates the unmatched subsequences 'P Q R S' and

'X Y Z' and gives them code symbols that show that they are alternatives in the same context. The model also creates an 'abstract' pattern that records the sequence of lower-level patterns in terms of their code symbols.

(2) Amongst the patterns that are generated in this way, some are 'good' in terms the principles of minimum length encoding and others are 'bad'. In the second stage of processing, the model measures the frequency with which each pattern may be recognised in the raw data and then it uses this information in a hill-climbing search amongst subsets of the Old patterns to find one or more sets of patterns that are good in terms of the principles of minimum length encoding. The remaining patterns may be discarded.

It is envisaged that, when the model is more fully developed, these two stages will be repeated so that the system can progressively bootstrap a set of patterns that are good in terms of the principles of minimum length encoding and represent a distillation of the patterns of redundancy in the original data.

If the potential of this model can be realised, the SP system should facilitate the automatic or semi-automatic construction of knowledge bases from raw medical data.

3.11 Classes and subclasses of diseases

One of the attractions of the SP system is that it allows concepts to be represented at multiple levels of abstraction (e.g., 'cat', 'mammal', 'vertebrate', 'animal') in the manner of object-oriented design and, via the building of multiple alignments, it allows a specific entity (such as "my cat Tibs") to be recognised at several different levels of abstraction [9,11].

To some extent, this idea is already illustrated by the example shown in Figure 3. The concept of 'fever', represented by the pattern in column 4 of the figure, may be seen as a superclass of all the diseases where the patient may be feverish. Likewise, the pattern for flu symptoms (column 3 in the figure) may be seen as a superclass of the diseases in which such symptoms may be seen.

By contrast with the classification of animals and plants, the hierarchy of diseases tends to be relatively flat. However, there is scope for the recognition of classes and subclasses in the variants of diseases such as influenza and diabetes. With the SP system, each variant of a given disease may be recorded as a pattern that specifies the symptoms that are characteristic of the variant. Provided that pattern contains a symbolic link to another pattern describing the main symptoms of the disease, there is no need to repeat those symptoms redundantly in each of the variants.

4 Comparison with alternatives

As mentioned in the introduction, a wide variety of philosophies and systems have been applied to the problem of medical diagnosis. In this section, I briefly review some of the more prominent of these approaches and compare them with the SP approach, as described in this paper.

4.1 Rule-based systems

Rule-based systems (like the well-known MYCIN system [21]) contain if-then rules where the 'if' side of any rule is a collection of one or more conditions for the rule to fire connected by logical operators such as 'AND', 'OR' (which may be inclusive or exclusive) and 'NOT'. By contrast, the SP system expresses all knowledge in the form of patterns.

At first sight, SP patterns lack the expressive power of if-then rules. But the effect of such rules can be modelled within the SP system if that is required [16,13]. And if medical diagnosis is viewed as a process of pattern recognition (as in this paper), then SP patterns and the SP framework are, arguably, a more natural and flexible medium for the representation and processing of knowledge than are if-then rules.

To illustrate this last point, we can express the distinctive features of influenza by a rule such as:

```
IF chills AND cough AND headache AND aching muscles AND runny nose AND sore throat
     THEN influenza (probability = 0.9)
```

Although there may be a probability associated with the rule (as shown), the rule has an intrinsic logic which, if strictly applied, means that the rule will *only* fire if all the conditions are satisfied. By contrast, a pattern like the one shown in column 3 of Figure 3 may appear in the best alignment when any reasonably large subset of its symbols have been matched.

If one attempted to achieve this kind of flexibility with an if-then rule using combinations of AND, OR and NOT, the rule would become very complex. Alternatively, one might split up the rule into a number of smaller rules, one for each symptom or combination of two or three symptoms—but again the result would be relatively complex.

4.1.1 Probabilities

In systems like MYCIN and some of its successors, the 'probabilities' that the system calculates are really measures of confidence without the theoretical under-

pinnings of probability theory. In other systems, "... formal approaches based on probability theory are precise but can be awkward and non-intuitive to use." [22, p. 272]. By contrast, the SP framework allows true probabilities to be calculated quite simply (see [11]) and strictly in accordance with established theory (as described in sources such as [23]).

4.2 Neural networks

One of the attractions of artificial neural networks for the support of medical diagnosis is that they can be trained with appropriate data, thus by-passing the need for the manual compilation of knowledge by medical experts or knowledge engineers. However, "A major drawback is that 'knowledge' embedded [in the neural network] is cryptically coded as a large number of weights and activation values. As a consequence, the lack of neural network validation tools is often one of the reasons limiting their use in practice, especially in the context of medical diagnosis where physicians cannot trust a system without explanation of its decisions." [24, pp. 141–142].

While there may be scope for extracting rules from a trained neural network (*ibid.*), this adds complexity and uncertainty to the technology and defeats the other main attraction of a neural network: as a classifier of specific cases in terms of the learned knowledge.

As a system for unsupervised learning of knowledge structures from raw data, the SP system is not yet a rival to existing neural network systems. However, the system has clear potential for unsupervised learning and, if that potential can be realised, the system has the advantage that its knowledge is stored in a form that can be read and understood by people. Meanwhile, if it is supplied with knowledge about diseases derived from experts or text books, it can be used for diagnostic classification of individual patients.

4.3 Fuzzy logic

Given the variability of diseases and other uncertainties associated with medical diagnosis (Section 3.6), fuzzy logic (see, for example, [25]) has the obvious attraction that it has been designed with the explicit intention of providing a model for 'fuzzy' concepts and 'fuzzy' operations on them.

In purely theoretical terms, the field of fuzzy logic may be criticised because it introduces a fairly elaborate conceptual framework to accommodate the undoubtedly fuzzy nature of many human concepts but this conceptual framework is poorly integrated with other ideas about the nature of human cognition. By contrast, the SP

theory grew out of research in psychology and it provides a plausible model for several aspects of human perception and cognition [9].

Considerations of that kind may be discounted as not relevant to the practicalities of medical diagnosis. But in that connection fuzzy logic has the drawback that it introduces another layer of complexity to the already difficult process of eliciting knowledge from medical experts [26]. There seems to be some scope for ameliorating this problem by the provision of appropriate tools (*ibid.*) but the basic problem remains. By contrast, the SP system allows concepts to be expressed in a simple, intuitive manner and, at the same time, it accommodates much, perhaps all, of the fuzziness of medical diagnosis.

4.4 Bayesian networks

Two of the main differences between Bayesian networks (see, for example, [20]) and the SP system are:

- Bayesian networks focus on the binary relationship between any given node in the network and each of its parent nodes (if any). In this respect, they inherit some of the thinking behind if-then rules. By contrast, the SP system is oriented towards the representation and processing of associations (expressed as patterns) that may contain arbitrarily many elements.
- Correspondingly, any given Bayesian network stores its statistical knowledge in the form of tables of conditional probabilities, one for each node in the network. By contrast, the SP system stores its statistical knowledge in the form of integers, one for each pattern, representing the absolute or relative frequency of the corresponding association in some domain.

These and related differences seem to underlie some of the apparent drawbacks of Bayesian networks compared with the SP framework:

- The directional nature of Bayesian networks does not sit easily with the non-directional nature of medical syndromes.
- The process of calculating probabilities of inferences in a Bayesian network is substantially more complicated than the calculation of probabilities for alignments and inferences in the SP framework.
- The tables of conditional probabilities required for Bayesian networks are significantly more complex than simple measures of frequency that are used in the SP system. Notwithstanding the development of special methods for eliciting conditional probabilities from experts [27], the process of building up the necessary tables of conditional probabilities is likely to be much harder than measuring or estimating an integer value for each disease, reflecting its absolute or relative frequency in a given domain.

4.5 Case-based reasoning

A major attraction of case-based reasoning in medical diagnosis (see, for example, [28,29]) is that, compared with many of the alternatives, it can considerably simplify the process of acquiring the necessary knowledge. In its simplest form, a case-based system merely requires a description of one or more specific examples of each disease and a search algorithm that can find exact matches or good partial matches between the symptoms of a given patient and one or more of the stored records.

In some respects, the SP system is like a case-based system and it could indeed be used like a case-based system. To use it in this way, each of the Old patterns should represent a specific case (including its diagnosis) and the New pattern or patterns should represent the symptoms of a patient for whom a diagnosis is required. The capabilities of the system for finding exact matches and good partial matches between patterns will allow it to retrieve patterns for previously-diagnosed cases that are similar to any given current case.

The main advantages of the SP system compared with the case-based approach to diagnosis are:

- It facilitates the description of diseases in generalised terms without the need to specify exact values for every category of symptom. In our main example, we saw how the temperature of a patient with a disease like influenza may be specified as a range of alternative values (Section 3.6). Any other category of symptom may be treated in the same way.
- It allows one to specify clusters of symptoms that are found in two or more different diseases and it allows one to describe diseases at two or more levels of abstraction (Section 3.11). Both of these things facilitate the description of diseases without the need to repeat information unnecessarily where similar patterns are found in different diseases or varieties of disease.

5 Conclusion

As we have seen, the SP system accommodates the main elements of medical diagnosis, viewed as a problem of pattern recognition, and there are reasons to believe that it may also provide support for causal reasoning in medical diagnosis. However, the SP62 model is a prototype that serves for research and demonstration: it is not yet a system with ‘industrial strength’. The main developments that are needed to reach that goal are:

- The provision of a well-designed graphical user interface.

- There is probably scope for improvements in the search methods that are used within the system.
- There is scope for the application of parallel processing both to improve the scaling properties of the system (Section 2.2.1) and to increase absolute speeds of processing.
- Naturally, the system needs to be provided with appropriate knowledge. For each area of application, a set of patterns needs to be developed that describes the diseases and symptom clusters in that domain.
- At some stage after the development of a realistic knowledge base, the performance of the system must be validated against the judgement of human medical experts.

The potential payoff from these developments is a system that allows knowledge about diseases to be expressed in a simple, intuitive manner, that can cope with errors and uncertainties in knowledge about diseases and knowledge about individual patients, that simplifies the acquisition and storage of statistical information, that calculates true probabilities of diagnoses, and smooths the path to the automatic or semi-automatic abstraction of medical knowledge in the future.

References

[1] F. Attneave, Some informational aspects of visual perception, Psychological Review 61 (1954) 183–193.

[2] H. B. Barlow, Trigger features, adaptation and economy of impulses, in: K. N. Leibovic (Ed.), Information Processes in the Nervous System, Springer, New York, 1969, pp. 209–230.

[3] J. G. Wolff, Learning syntax and meanings through optimization and distributional analysis, in: Y. Levy, I. M. Schlesinger, M. D. S. Braine (Eds.), Categories and Processes in Language Acquisition, Lawrence Erlbaum, Hillsdale, NJ, 1988, pp. 179–215, copy: www.cognitionresearch.org.uk/lang_learn.html#wolff_1988.

[4] N. Chater, Reconciling simplicity and likelihood principles in perceptual organisation, Psychological Review 103 (3) (1996) 566–581.

[5] R. J. Solomonoff, A formal theory of inductive inference. parts I and II, Information and Control 7 (1964) 1–22 and 224–254.

[6] C. S. Wallace, D. M. Boulton, An information measure for classification, Computer Journal 11 (2) (1968) 185–195.

[7] J. Rissanen, Modelling by the shortest data description, Automatica-J, IFAC 14 (1978) 465–471.

[8] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag, New York, 1997.

[9] J. G. Wolff, Information compression by multiple alignment, unification and search as a unifying principle in computing and cognition, Artificial Intelligence Review 19 (3) (2003) 193–230, copy: http://arxiv.org/abs/cs.AI/0307025.

[10] J. G. Wolff, Information compression by multiple alignment, unification and search as a framework for human-like reasoning, Logic Journal of the IGPL 9 (1) (2001) 205–222, first published in the *Proceedings of the International Conference on Formal and Applied Practical Reasoning (FAPR 2000)*, September 2000, ISSN 1469–4166. Copy: www.cognitionresearch.org.uk/papers/pr/pr.htm.

[11] J. G. Wolff, Probabilistic reasoning as information compression by multiple alignment, unification and search: an introduction and overview, Journal of Universal Computer Science 5 (7) (1999) 418–462, copy: http://arxiv.org/abs/cs.AI/0307010.

[12] J. G. Wolff, Syntax, parsing and production of natural language in a framework of information compression by multiple alignment, unification and search, Journal of Universal Computer Science 6 (8) (2000) 781–829, copy: http://arxiv.org/abs/cs.AI/0307014.

[13] J. G. Wolff, Mathematics and logic as information compression by multiple alignment, unification and search, Tech. rep., CognitionResearch.org.uk, copy: http://arxiv.org/abs/math.GM/0308153. (2002).

[14] J. G. Wolff, Unsupervised grammar induction in a framework of information compression by multiple alignment, unification and search, in: C. de la Higuera, P. Adriaans, M. van Zaanen, J. Oncina (Eds.), Proceedings of the Workshop and Tutorial on Learning Context-Free Grammars, 2003, pp. 113–124, this workshop was held in association with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2003), September 2003, Cavtat-Dubrovnik, Croata. Copy: http://arxiv.org/abs/cs.AI/0311045.

[15] J. G. Wolff, Unsupervised learning in a framework of information compression by multiple alignment, unification and search, Tech. rep., CognitionResearch.org.uk, copy: http://arxiv.org/abs/cs.AI/0302015. (2002).

[16] J. G. Wolff, 'Computing' as information compression by multiple alignment, unification and search, Journal of Universal Computer Science 5 (11) (1999) 777–815, copy: http://arxiv.org/abs/cs.AI/0307013.

[17] A. Church, The Calculi of Lamda-Conversion, Vol. 6 of Annals of Mathematical Studies, Princeton University Press, Princeton, 1941.

[18] E. L. Post, Formal reductions of the general combinatorial decision problem, American Journal of Mathematics 65 (1943) 197–268.

[19] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, Extensible Markup Language (XML) 1.0 (second edition), Tech. rep., World Wide Web Consortium, W3C recommendation, 6 October 2000. Copy: www.w3.org/TR/REC-xml (2000).

[20] J. Pearl, Probabilistic Reasoning in Intelligent Systems, revised second printing Edition, Morgan Kaufmann, San Francisco, 1997.

[21] E. H. Shortliffe, Computer-Based Medical Consultations: MYCIN, Elsevier/North Holland, New York, 1976.

[22] J. Fox, D. Glasspool, J. Bury, Quantitative and qualitative approaches to reasoning under uncertainty in medical decision making, in: S. Quaglini, P. Barahona, S. Andreassen (Eds.), Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001, Vol. 2101 of Lecture Notes in Computer Science, Springer, 2001, pp. 272–282.

[23] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley, New York, 1991.

[24] G. Bologna, A model for single and multiple knowledge based networks, Artificial Intelligence in Medicine 28 (2003) 141–163.

[25] G. J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentic Hall PTR, Upper Saddle River, NJ, 1995.

[26] K. Boegl, K.-P. Adlassnig, Y. Hayashi, T. E. Rothenfluh, H. Leitich, Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system, Artificial Intelligence in Medicine 30 (2004) 1–26.

[27] L. C. van der Gaag, S. Renooij, C. Witteman, B. Aleman, B. Taal, Probabilities for a probabilistic network: a case study in oesophageal cancer, Artificial Intelligence in Medicine 25 (2002) 123–148.

[28] R. Schmidt, L. Gierl, Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes, Artificial Intelligence in Medicine 23 (2001) 171–185.

[29] K.-D. Althoff, R. Bergmann, S. Wess, M. Manago, E. Auriol, O. I. Larichev, A. Bolotov, Y. I. Zhuravlev, S. I. Gurov, Case-based reasoning for medical decision support tasks: the Inreca approach, Artificial Intelligence in Medicine 12 (1998) 25–41.