From AAAI Technical Report FS-98-04. Compilation copyright © 1998, AAAI. (www.aaai.org). All rights reserved.

# Diagrammatic Reasoning and Color

**Michael Anderson & Chris Armen**

Department of Computer Science
University of Hartford
200 Bloomfield Avenue
West Hartford, Connecticut USA 06117
anderson@hartford.edu, armen@hartford.edu

**Abstract**

The theory of inter-diagrammatic reasoning is extended to include the use of color values in diagrams. The syntax of this more general theory, its subsumption of the original theory, and the ramifications of this extension are presented. An example application of this extension in the domain of DNA sequence assembly is described.

## Introduction

The theory of *inter-diagrammatic reasoning* (Anderson and McCartney 1995; Anderson and McCartney 1996; Anderson and McCartney 1997) defines an approach to computation with diagrammatic representations that strives for generality. It provides a simple diagram syntax and a general set of operators that can be used to reason with collections of related diagrams in a wide range of domains. Example applications include, among others, heuristic development in game playing, learning from musical notation, and inference with cartograms.

This work has, in part, been motivated by a desire to explore the boundaries of what can be computed via diagrammatic representations alone. It may be the case that these representations will need to be augmented by others representations to fully represent many problem domains. We believe that exploring the limits of diagrammatic representations is necessary to bring our understanding of them on par with other, better understood representation schemes. Only when parity is achieved in our understanding of all representation schemes can we make informed judgments concerning their uses.

Another factor motivating this work is our interest in investigating general approaches to computing with diagrammatic representations that prove useful across wide varieties of domains. Although any one such approach is not likely to be all-inclusive, development of a number of such approaches can only serve to illuminate the dimensions of the problem. For instance, by partitioning domains into classes of those that can and cannot be serviced by a given approach, defining properties of these classes can be more readily deduced.

In the spirit of these motivations, we have extended the theory of inter-diagrammatic reasoning to include the use of color values in diagrams. Generally acknowledged as an interesting dimension of diagrammatic representations, color has not been formalized for use in diagrammatic reasoning. We present a formalism based on a *subtractive CMY color system* (Wyszecki and Stiles 1982) that both subsumes the original inter-diagrammatic theory and provides new dimensions to diagrammatic computation.

In the following we briefly recap salient points of the theory of inter-diagrammatic reasoning, present the syntax of the more general theory generated by its extension to color, and show the effect of this extension on inter-diagrammatic operators. The newly available color dimensions are codified and some possible semantic interpretations of these dimensions are explored. Next, an example application of the extended theory in the domain of DNA sequence assembly is presented. This is followed by a brief discussion of related work and, finally, we offer our conclusions.

## A More General Theory of Inter-Diagrammatic Reasoning

We generalize the theory of inter-diagrammatic reasoning to include the use of color in a diagram. As in previous incarnations of the theory, the assumption is that, most generally, one can syntactically define a *diagram* to be a tessellation (tiling) of a planar area such that it is completely covered by atomic two dimensional regions or tesserae (tiles). Such a definition is broad enough to include arbitrarily small tesserae (points, at the limit), pixels, and, at the other end of the spectrum, domain-specific entities such as countries, regions, blocks, etc.

Previously, the theory assumed tesserae were only to take their values $v$ from an $l$ valued, additive gray scale (denoted $\{v_0, \ldots, v_{l-1}\}$) increasing from a minimum value $v_0$, WHITE, to a maximum value $v_{l-1}$, BLACK. Intuitively, these gray scale values correspond to a discrete set of

| Saturation \ Hue | Equal ($i=j=k$) | Exclusive ($i=j=0, i=k=0$, or $j=k=0$) | Unconstrained |
|---|---|---|---|
| Two-valued ($I=J=K=2$) | *Black & White* $\{v_0, v_{I-1}\}$ | *Primary & White* $\{v_{0,0,0}, v_{I-1,0,0}, v_{0,J-1,0}, v_{0,0,K-1}\}$ | *CMY & White* $\{v_{0,0,0}, v_{I-1,0,0}, v_{I-1,J-1,0}, v_{I-1,0,K-1}, v_{0,J-1,0}, v_{0,J-1,K-1}, v_{0,0,K-1}, v_{I-1,0,0}, v_{I-1,J-1,K-1}\}$ |
| Multi-valued ($I=J=K=c$) | *Gray Scale* $\{v_0, \ldots, v_{I-1}\}$ | *Primary Scale* $\{v_{0,0,0}, v_{1,0,0}, \ldots, v_{I-1,0,0}, v_{0,1,0}, \ldots, v_{0,J-1,0}, v_{0,0,1}, \ldots, v_{0,0,K-1}\}$ | *CMY Gamut* $\{v_{0,0,0}, \ldots, v_{I-1,J-1,K-1}\}$ |

*Table 1*: Saturation and hue color dimensions

transparent gray filters that, when overlaid, combine to create a darker filter to a maximum of BLACK.

### Extension to Color

The generalization of the theory permits tesserae to take their values from an $I$, $J$, $K$ valued subtractive CMY color scale. Intuitively, these CMY (Cyan, Magenta, Yellow) color scale values (denoted $v_{i,j,k}$) correspond to a discrete set of transparent color filters where $i$ is the cyan contribution to a filter's color, $j$ is the magenta contribution, and $k$ is the yellow contribution. When overlaid, these filters combine to create other color filters from a minimum of WHITE ($v_{0,0,0}$) to a maximum of BLACK ($v_{I-1,J-1,K-1}$).

The following binary operators, modified to deal with color values, take two diagrams, $d_1$ and $d_2$, of equal dimension and tessellation and return a new diagram where each tessera has a value that is some function of the values of the two corresponding tesserae, $v_{i1,j1,k1}$ and $v_{i2,j2,k2}$, in the operands:

*OR*, denoted $d_1 \vee d_2$, returns the *maximum* of each pair of tesserae where the maximum of two corresponding tesserae is defined as

$v_{max(i1,i2),\ max(j1,j2),\ max(k1,k2)}$

*AND*, denoted $d_1 \wedge d_2$, returns the *minimum* of each pair of tesserae where the minimum of two corresponding tesserae is defined as

$v_{min(i1,i2),\ min(j1,j2),\ min(k1,k2)}$

*OVERLAY*, denoted $d_1 + d_2$, returns the *sum* of each pair of tesserae where the sum of values of corresponding tesserae is defined as

$v_{min(i1+i2,\ I-1),\ min(j1+j2,\ J-1),\ min(k1+k2,\ K-1)}$

*PEEL*, denoted $d_1 - d_2$, returns the *difference* of each pair of tesserae the difference of values of corresponding tesserae is defined as

$v_{max(i1-i2,\ 0),\ max(j1-j2,\ 0),\ max(k1-k2,\ 0)}$.

The following presents some of the ramifications of extending the theory to permit color values.

## Color Dimensions

Just as we find ourselves reaching for a colored marker to illustrate some dimension of information beyond the scope of our black one, we have extended the theory of inter-diagrammatic reasoning to encompass color values to provide us with new dimensions with which to represent information. Table 1 depicts the relationship between two defining dimensions of these color values, *saturation* and *hue*, and the sets of values that are defined by these dimensions. Saturation refers to the intensity of a given color, varying from WHITE (minimal saturation) to pure (maximal saturation). Hue refers to the number and relationship of primary colors that contribute to a given color. The saturation dimension is permitted two states: *Two-valued*, where saturation can be only minimal and maximal ($I=J=K=2$), and *Multi-valued*, where saturation can vary from minimal to maximal ($I=J=K=c$). The hue dimension is permitted three states: *Equal* where $i = j = k$, *Exclusive* where $i = j = 0$, $i = k = 0$, or $j = k = 0$, and *Unconstrained* where $i$, $j$, $k$ can vary at will.

The current theory subsumes the previous theory in that black & white and gray scale values previously used are proper subsets of the values used by this more general theory. Table 1 shows that they are related in that both are defined by *Equal* hue. When saturation is *Two-valued*, black & white values arise. When saturation is *Multi-valued*, gray scale values emerge. Letting $v_i$ denote $v_{i,j,k}$ in these sets resolves the previous syntax.

*Exclusive* hue permits any one subscript to contribute to a color value while holding the other two to 0. When combined with *Two-valued* saturation, a set of four colors

(WHITE, cyan, magenta, and yellow) is produced. When combined with *Multi-valued* saturation, a set of three color ranges is produced, each range comprised of a primary color that varies from minimal saturation to maximal saturation. One such range, for example, is produced by holding $j$ and $k$ to zero while varying $i$ from $0$ to $I-1$ will produce a set of color values that varies from WHITE (at $0$) to pure cyan.

*Unconstrained* hue permits all subscripts to contribute to a color value. When combined with *Two-valued* saturation, a set of eight colors (WHITE, cyan, magenta, yellow, red, green, blue, and BLACK) is produced. When combined with *Multi-valued* saturation, a color set is produced that is bounded only by the *gamut* or range of the CMY color scale, a subrange of the gamut of the visible spectrum.

The following is an example of how these new color dimensions might be semantically interpreted.

## An Example: Scheduling a Meeting

A simple variant of the *scheduling problem* arises when a number of people need to arrange a common meeting time. An acceptable range of possible meeting times is chosen, schedules are compared to see when common free time exists, and, finally, a meeting time is decided. This process can be accomplished diagrammatically given a suite of diagrams representing the collected schedules of those attempting to arrange a meeting time. A diagram is comprised of tesserae for each possible meeting time in the chosen range. Values these tesserae take, and the possible interpretations of these values, are dependent upon the color dimension values chosen. The following explores the use of various saturation and hue dimensions combinations in this scheduling problem.

Choosing two-valued saturation and equal hue color dimensions, constrain tesserae to take BLACK and WHITE values only. A semantic mapping of BLACK to *unavailable* and WHITE to *available* permits interpretation of a diagram as a schedule. Individual schedules ($d_1$, $d_2$, ..., $d_N$) are created as diagrams of BLACK and WHITE tesserae. When all such schedules are ORed together,

$$(d_1 \vee d_2 \vee \ldots \vee d_N),$$

only corresponding tesserae that have WHITE values in a diagrams will remain WHITE in the final ORed diagram. These, then, will represent free time common to all concerned and all that remains to be done is to choose one for a meeting time.

A more informative schedule can be developed if *Multi-valued* saturation and *Equal* hue dimensions are chosen. The gray scale values derived can be used to represent a wider range of availability than can BLACK and WHITE values alone. For example, some gray value between BLACK and WHITE can be used to represent a level of availability somewhere between *available* and *unavailable*, say *semi-available*. Semi-availability provides a way to mark a meeting time as possible but less desirable than a completely available time. In this example, individual schedules are created as diagrams comprised of three-valued tesserae: BLACK, WHITE, and gray. Figure 1a shows three such diagrams with rows corresponding to the days Monday through Friday and the columns representing the hours 9:00 to 5:00.

Since ORing diagrams is defined as taking the maximum value of corresponding tesserae, ORing together the schedules of all concerned will produce a diagram with values such that:

WHITE represents time available to all,
BLACK represents time unavailable to at least one, and
gray represents time semi-available to at least one with others being either available or semi-available (i.e. no one is unavailable).

A simple process is used to find the best time for all to meet: if one or more WHITE tesserae exist, choose one, otherwise choose a gray tessera.

Unfortunately, ORing schedules provides no way to differentiate between semi-available times comprised of different numbers of semi-available people. For instance, time semi-available to one and available to all others is represented by the same gray value that represents time only semi-available to all. If the gray scale value $i$ used to represent semi-availability is chosen such that, given $n$ schedules, $n \times i < I-1$ (i.e. the maximum OVERLAYed value is guaranteed to be less than BLACK), such a differentiation can be made. With this constraint, OVERLAYing all schedule diagrams,

$$(d_1 + d_2 + \ldots + d_N),$$

produces a variety of gray scale values ranging from WHITE (all are available) to BLACK (at least one is unavailable). The intensities of the gray values in between correspond to the number of those who are semi-available — the darker the gray, the more who are only semi-available. The process of determining the best meeting time now consists of choosing a WHITE tessera if present and, if not, choosing the lightest gray tessera available. Such a gray tessera will correspond to the most desirable of the semi-available meeting times.

Allowing tesserae to take values from the full CMY gamut (*Multi-valued* saturation and *Unconstrained* hue dimensions) provides a richer set of possible semantic interpretations than those previously available. One such interpretation correlates primary colors with people. For example, assume there are three people trying to schedule a mutually agreeable meeting time. Each has a schedule in gray scale as previously specified ($d_1$, $d_2$, $d_3$) with BLACK, gray, and WHITE corresponding to unavailability,

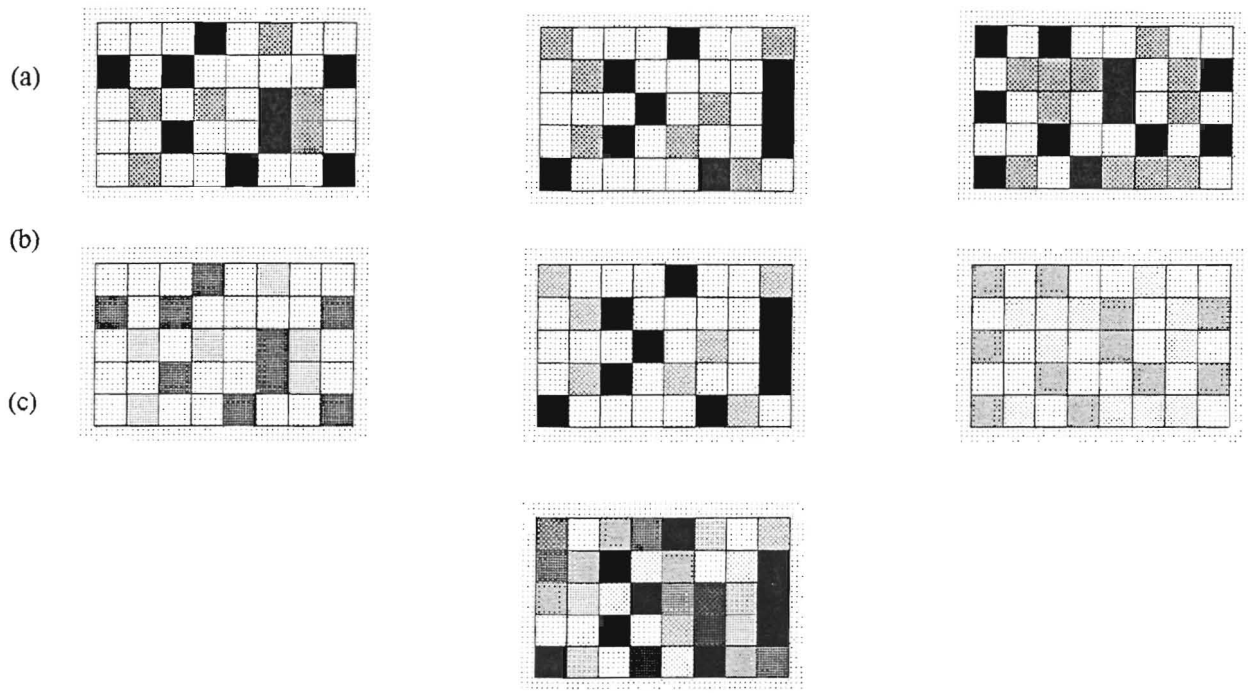


*Figure 1*: Scheduling meeting times
(a) $d_1, d_2, d_3$ (b) $d_1 \wedge d_C, d_2 \wedge d_M, d_3 \wedge d_Y$ (c) $(d_1 \wedge d_C) + (d_2 \wedge d_M) + (d_3 \wedge d_Y)$

semi-availability, and availability (see Figure 1a). ANDing one of these diagrams with a diagram completely covered in a given primary color results in a diagram that represents unavailability, semi-availability, and availability as intensities of that primary color (see Figure 1b). For example,

ANDing a fully saturated cyan-valued tessera ($v_{i-1,0,0}$) with a black-valued tessera ($v_{i-1, j-1, k-1}$) produces a fully saturated cyan-valued tessera ($v_{i-1,0,0}$),

ANDing a fully saturated cyan-valued tessera ($v_{i-1,0,0}$) with a white-valued tessera ($v_{0,0,0}$) produces a white-valued tessera ($v_{0,0,0}$), and

ANDing a fully saturated cyan-valued tessera ($v_{i-1,0,0}$) with an arbitrary gray-valued tessera ($v_{i,j,k}$) produces a cyan-valued tessera ($v_{i,0,0}$) whose saturation is dependent upon the value of $i$.

In the example, each of the three schedule diagrams is ANDed with one of the three primary color diagrams ($d_C$, $d_M$, $d_Y$) and three new schedule diagrams result: a cyan schedule diagram, a magenta schedule diagram, and a yellow schedule diagram. When these diagrams are OVERLAYed,

$$(d_1 \wedge d_C) + (d_2 \wedge d_M) + (d_3 \wedge d_Y),$$

the resulting diagram is produced with a wide variety of values (see Figure 1c), each with a particular domain specific meaning:

WHITE represents time available to all,
BLACK represents time unavailable to all,
gray represents time semi-available to all,
fully saturated CMY (cyan, magenta, or yellow) represent time unavailable to one particular person and available to all others,
$i$-saturated CMY represent time semi-available to one particular person and available to all others,
fully saturated RGB (red, green, or blue) represent time unavailable for two particular people and available to the other,
$i$-saturated RGB represent time semi-available for two particular people and available to the other, and
three less easily described color ranges represent 1) time unavailable to one particular person and semi-available to all others, 2) time unavailable to two particular people and semi-available to the other, and 3) time unavailable to one particular person, semi-available to one particular person, and available to one particular person.

Given this wide range of possible values, solutions to a wider variety of problems are implicit in this diagram. First, the solution to the problem as originally stated can be found by choosing a WHITE tessera if it exists or, if not, choosing a gray, $i$-saturated CMY, or $i$-saturated RGB tessera instead. Gray, $i$-saturated CMY, or $i$-saturated

RGB values correlate to gray scale intensities in the previous solution in that they correspond to the least to the most desirable of the semi-available meeting times, respectively.

Since individual schedules are preserved within the resulting diagram via color, information pertaining to individuals, e.g. rank, can be taken into consideration before a meeting time is settled on. For example, if no WHITE tessera exists, an *i*-saturated RGB tessera where the highest ranking individual is available might represent a more desirable meeting time than an *i*-saturated CMY tessera where the highest ranking individual is only semi-available. Clearly, this is only one example of an entire class of such problems.

It is also possible to determine by inspection contiguous blocks of time that are not *unavailable* to anyone. Since these tesserae, by definition, will not include any fully saturated color, their saturation level will be at most *i* and will appear lighter than those tesserae that include fully saturated color. This property provides solutions to problems that require a number of contiguous meeting times either on the same day or at the same time every day.

The extension of the theory of inter-diagrammatic reasoning to include color values provides a richer set of possible semantic interpretations than the theory previously encompassed. For example, gray scale alone might be useful for representing a quantity or quality, whereas color values could simultaneously represent the notions of quantity or quality by saturation and the notions of category or class by hue.

A description of a real world domain in which this extension can be applied follows.

## An Application: DNA Sequence Assembly

Much of the growing field of computational molecular biology consists essentially of problems about character strings. One strand of a double-stranded DNA molecule can be viewed as a string drawn from a four-character alphabet $\Sigma=\{a,c,g,t\}$. Each of these characters represents a single *base*; a molecule of human DNA is about $10^8$ bases in length. The pairs $\{a,t\}$ and $\{c,g\}$ are complementary; in the structure of the DNA molecule, the bond between each complementary pair constitutes a "rung" on the twisted ladder formed by the two *strands* and these bonds. Thus, the bases of either of the two strands completely determine the sequence of the opposing strand.

*DNA sequencing* is the process of determining the bases in a particular molecule of DNA The portion of the molecule of interest might contain 100,000 bases, but current technology allows a sequence of only several hundred bases to be determined directly. We will refer to the laboratory process of determining the bases of a relatively short sequence as *reading* the sequence. A common laboratory protocol to deal with this problem, called *shotgun sequencing*, is to create several copies of the region of DNA of interest. These copies are then fragmented in what amounts to a random fashion; the process of fragmentation can be controlled sufficiently to guarantee that many fragments will be of a length that can be read in the laboratory. The laboratory data thus consists of a set of character strings of various lengths, each of which is drawn from some portion of the original sequence. Due to the random nature of the fragmentation, these strings are likely to have a high degree of overlap. The *sequence assembly* problem is the computational task of reconstructing the original sequence from the overlapping strings. (For a more complete discussion of shotgun sequencing, see (Myers 1997).)

As stated above, the problem of finding overlapping fragments is an *exact string matching* problem, in which we seek a suffix of one string and a prefix of another string that are identical. Unfortunately, errors can occur while reading the sequence of bases in the lab, with the result that the data do not exactly match the underlying string, and overlapping fragments do not exactly match each other. Thus determining overlaps is an instance of the more difficult problem of *approximate string matching*. There are two types of errors that can occur. A *substitution error* occurs when one base is mistaken for another and therefore substituted for it in the data. An *insertion/deletion error*, or *indel*, occurs when a base is not read, or when a nonexistent base is inserted into the data.

An additional complication arises from the fact that the fragments originate from both strands of the DNA molecule. Due to the structure of DNA, some of the overlaps will therefore be detected only as matches between one character string and the *reverse complement* of the other, where *complement* refers to the $\{a,t\}$ and $\{c,g\}$ pairings.

Numerous algorithms exist for automating the process of sequence assembly (Peltola *et al.* 1984; Blum *et al.* 1994; Kececioglu and Meyers 1995; Armen and Stein 1998). These algorithms are generally effective at coalescing the initial set of strings down to a relatively small set of longer strings. At this point, however, the strings may be able to be ordered in several ways, with no clear-cut best choice that an algorithm can identify. This last stage of assembling the remaining strings is left to the biologist as a manual task. The job is visual and heuristic in nature; i.e. the biologist is searching for "likely" overlaps, in the presence of substitution errors and indels. The final choices depend on a wide range of domain knowledge and experience brought to the task by the biologist. We propose a diagrammatic computation of the

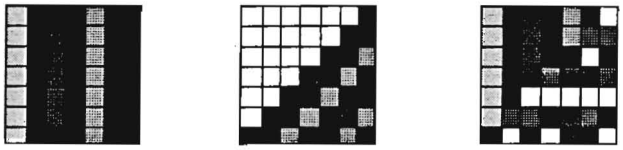


*Figure 2*: Exact string matching
(a) $d_L$ (b) $d_R$ (c) $d_L \otimes d_R$
(The matching alignment appears in row 5 as all WHITE ($v_{0,0,0}$) tesserae.)

exact and approximate matches in order to expedite this manual task.

## Exact String Matching

We first present a diagrammatic solution to the problem of determining the overlap between two strings in the presence of substitution errors only. Given two strings L and R, and a maximum possible overlap $m$, our approach allows the biologist to simultaneously view all $m$ possible overlaps between L and R in which L appears to the left of R. We describe the diagrammatic representation and the computation required.

We use a system with *Two-valued* saturation and *Unconstrained* hue; recall from Table 1 that this system allows us eight colors. Of these we will use six; four are for the bases, and BLACK and WHITE will have special significance. In choosing our base colors, we have adhered as much as possible to the conventions that have evolved in biology; unfortunately, we cannot do so entirely, because we want to represent complementary bases as complementary colors in the CMY system. (Our reason for this is discussed below.) We let $I=J=K=2$, and assign our colors as follows:

$t$ = red = $v_{0,1,1}$ ,
$a$ = cyan = $v_{1,0,0}$ ,
$c$ = blue = $v_{1,1,0}$ ,
and $g$ = yellow = $v_{0,0,1}$ .

In addition to the primitive operators described above, we shall make extensive use of the XOR ($\otimes$) operation, which we define as follows using them. As before, let $d_1$ and $d_2$ be diagrams with equal dimension and tessellation. Then we define $d_1 \otimes d_2$ to be the diagram that results from

$$(d_1 \vee d_2) - (d_1 \wedge d_2).$$

We observe that if two corresponding tessera have the same color, the result of the XOR will be $v_{0,0,0}$ (WHITE), and if the corresponding tessera are complementary, the result of the XOR will be $v_{1,1,1}$ (BLACK).

We refer to the diagrams corresponding to L and R as $d_L$ and $d_R$ , respectively. Both consist of an $m \times m$ grid of tiles, as in Figure 2. Each row of $d_L$ is identical, and consists of the colors corresponding to the last $m$ bases of L. The diagram $d_R$ is built as follows. The top row consists of the color corresponding to the first base in R as the rightmost tile; the remaining tiles in the top row are WHITE ($v_{0,0,0}$). The second row has the colors for the first two bases of R as the rightmost two tiles; the remaining tiles are WHITE. The pattern continues in this manner through the bottom row, the tiles of which correspond to the prefix of R of length $m$.

We now compute $d_L \otimes d_R$. The upper-left half of the result looks just like $d_L$, as WHITE ($v_{0,0,0}$) and XOR merely performs an identity operation. The area of interest is the bottom-right half of the diagram. An exact match of length $n$ will be indicated by a complete row of WHITE in the $n$th row, from the right edge of the diagram through the diagonal element. Approximate matches in which substitution errors have occurred will be seen as rows that are almost all WHITE, with a mismatch being seen as some color. (For instance, a $c$ matched with a $t$ would result in $v_{1,0,1}$, or green.) The biologist can determine with a simple visual inspection whether there are any likely overlaps, based on the frequency and distribution of any substitution mismatches.

The foregoing assumes that the overlapping strings originated from two copies of the same strand of the DNA molecule. In fact, in some cases they might be derived from opposing strands, and therefore as described above, one string might match the reverse complement of the

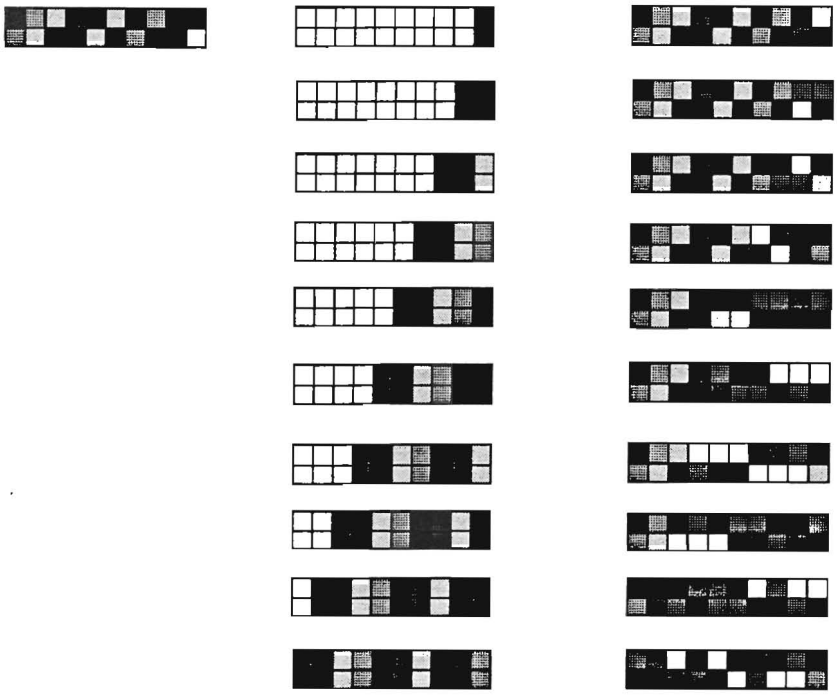

*Figure 3*: String matching with indels.
(a) $D_L$ (b) $D_R(n)$ (c) $D_L \otimes D_R(n)$
(The matching alignment appears in row 7 as contiguous groups of WHITE ($v_{0,0,0}$) tesserae.)

other. Our system readily handles this case: $d_L$ can be used as it was originally constructed, and $d_R$ is built from R reversed, the diagrams XORed as before, and a match will be detected as a row of $v_{1,1,1}$ (BLACK).

## String Matching with Indels

We have shown how to detect overlap between two fragments in the presence of substitution errors, and how to handle the complication of reverse complement matches. However, this is a simplification of the real problem, in which indels can occur. Observe that a single indel can make a long section of overlap appear to be a mismatch:

```
...tgtacgtcaca
    \ \ / / / /
     ac tcacat..
```

We will first show how to determine overlap when there is a net gain (or loss) of one base in one of the strings. We then will describe how to generalize this approach to a gain or loss of $h$ indels for any constant $h$.

Our diagrams for this application are grids of $2 \times m$ tiles, where $m$ is once again the maximum anticipated overlap. Once again a single diagram $d_L$ corresponds to L. Its first row consists of the colors corresponding to the suffix of L of length $m$ and its second row consists of the colors corresponding to the suffix of L of length $m-1$, with a WHITE tile at the right. (See Figure 3a.) We define a suite of diagrams $D_R = \{d_R(1), d_R(2), \ldots, d_R(m)\}$, to represent R. For each $n$, $1 \le n \le m$, $d_R(n)$ consists of two identical rows. The rightmost $n$ tiles of each row correspond to the bases of the prefix of R of length $n$. The remaining $m-n$ tiles are WHITE. (See Figure 3b.)

For each diagram $d_R(n)$ in $D_R$, we compute $d_L$ XOR $d_R(n)$. Each iteration tests a pair of adjacent alignments of L and R, and a good match with a single indel will appear as a series of WHITE tiles, which may be split between the two rows due to indels. (See Figure 3c; the approximate match occurs with $d_R(7)$) As in the case of our first construction, simply reversing one of the input strings allows us to look for a reverse complement match corresponding to the case of fragments from opposite strands.

This approach generalizes to handle $h$ indels for any constant $h < m$. Instead of diagrams of dimension $2 \times m$, we require diagrams of dimension $(h+1) \times m$. The construction of $d_L$ and $D_R$ is as above, extended over the additional

rows. A "match" will now consist of series of horizontal WHITE bands of tesserae, each of which may occupy any of the $h+1$ rows. Each segment of the match is likely to be on an adjacent row to the previous, as most indels will usually consist of a single base.

We have shown how to solve problems in exact and approximate string matching using diagrams, in the particular domain of DNA sequence assembly. Although many symbolically-based algorithms exist for variations on these problems, we propose that the diagrammatic approach is particularly well suited to an application such as this in which the computation and human direction are so closely intertwined.

## Related Work

Research in diagrammatic reasoning is just beginning to flourish after a long dormancy it experienced being virtually abandoned after a brief flirtation in the early days of AI (Gelernter 1959; Evans 1962). See, for instance, (Larkin and Simon 1987; Narayanan and Chandrasekaran 1991; Narayanan 1992; Chandrasekaran, Narayanan, and Iwasaki 1993; Narayanan 1993; Glasgow 1993; Glasgow, Narayanan, and Chandrasekaran 1995; Anderson 1997) for a representative sample of this work. Inter-diagrammatic reasoning has been previously proposed as one way of using diagrammatic representations to solve problems (Anderson and McCartney 1995; Anderson and McCartney 1996; Anderson and McCartney 1997). The earliest work in diagrammatic reasoning can be considered the first example of inter-diagrammatic reasoning as well (Evans 1962). (Furnas 1992) presents interesting work that, like ours, explores the boundaries of computing with diagrammatic representations alone.

Work in color theory is well-established and provides a solid foundation upon which to develop approaches for dealing with color diagrammatic representations. The subtractive CMY color system is only one of many such systems. Others that might prove useful include the additive RGB color system as used by RGB monitors and the Hue, Saturation and Luminance system that portends to be a better mapping to how humans manipulate color. (Wyszecki and Stiles 1982) is a good overview. We believe we are among the first to use color to the benefit of diagrammatic reasoning.

Most algorithmic work on DNA sequencing has focused, naturally enough, on the part that can be fully automated. (Peltola *et al.* 1984) was one of the earliest such systems; (Kececioglu and Meyers 1995) presents a variety of more sophisticated modern approaches. One frequently-used heuristic is to seek an ordering of the overlapping fragments that minimizes length (or equivalently, that maximizes overlap); (Blum *et al.* 1994) and (Armen and Stein 1998) address this problem. (Meyers 1997) provides a thorough introduction to the computational problem in this area and their biological basis.

## Conclusion

In this work, we have extended the theory of inter-diagrammatic reasoning to incorporate color and explored how the new dimensions this extension provides might prove useful. This, in turn, has widened the scope of the theory to include problem domains not previously encompassed. Using this extension, we have extended the boundary of computing with diagrams into domains not previously explored including scheduling and DNA sequencing.

Clearly, this work does not exhaust the possibilities that the introduction of color presents. The added dimensions color brings are open to a wide variety of semantic interpretations. We envision future application of the theory to domains such as color cartograms, image databases, data mining, and others.

## References


Anderson, M., ed., 1997. Reasoning with Diagrammatic Representations II, Proceedings of the AAAI 1997 Fall Symposium, AAAI Press, MIT.

Anderson, M. and McCartney, R. 1995. Inter-diagrammatic Reasoning. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.

Anderson, M. and McCartney, R. 1996. Diagrammatic Reasoning and Cases. Proceedings of the 13th National Conference on Artificial Intelligence.

Anderson, M. and McCartney, R 1997. Learning from Diagrams. *International Journal of Machine Graphics & Vision*, Vol. 6 , No. 1, pp. 57-76.

Armen, C. and Stein, C., A 2-2/3 superstring approximation algorithm. *Discrete Applied Mathematics*, to appear.

Blum, A., Jiang, T., Li, M., Tromp, J., and Yannakakis, M. 1994. Linear approximation of shortest superstrings. *Journal of the ACM*, 41(4):630-647, July.

Chandrasekaran, B., Narayanan, N. and Iwasaki, Y. 1993. Reasoning with Diagrammatic Representations, *AI Magazine*, 14(2).

Evans, T.G. 1962. A Heuristic Program to Solve Geometry Analogy Problems. MIT AI Memo 46. (also in *Semantic Information Processing* as "A Program for the Solution of a Class of Geometric-analogy Intelligence-test Questions", 271-353, Minsky, M. L., ed. MIT Press, 1968).

Furnas, G. 1992. "Reasoning with Diagrams Only" pp. 118-123 in (Narayanan 1992)

Gelernter, H. 1959. Realization of a Geometry Theorem Proving Machine. Proceedings of an International Conference on Information Processing, 273-282. UNESCO House.

Glasgow, J. 1993. The Imagery Debate Revisited: A Computational Perspective in (Narayanan 1993).

Glasgow, J., Narayanan, N., and Chandrasekaran, B. 1995. *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, AAAI Press.

Kececioglu, J. and Myers, E.. Exact and approximate algorithms for the sequence reconstruction problem. *Algorithmica* 13:7, 1995.

Larkin, J. and Simon, H. 1987. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11, 65-99.

Myers, E. Is whole human genome sequencing feasible?, *Computational Methods in Genome Research*, (S. Suhai, ed.) Plenum Press (New York, 1997), 73-89.

Narayanan, N., ed. 1992. Working Notes of AAAI Spring Symposium on Reasoning with Diagrammatic Representations.

Narayanan, N., ed. 1993 Taking Issue/Forum: The Imagery Debate Revisited. *Computational Intelligence*, 9(4).

Peltola, H., Soderlund, H. and Ukkonen, E.. DEQUAID: a DNA sequence assembly program based on a mathematical model. *Nucleic Acids Research*,12:307-321, 1984.

Wyszecki, G. and Stiles, W. S. 1982. *Color Science*, 2nd ed., John Wiley and Sons.