Mathematical Geology, Vol. 19, No. 2, 1987

The Match Game: New Stratigraphic Correlation Algorithms[1]

Michael S. Waterman[2] and Robert Raymond, Jr.[3]

New algorithms for automatic correlation of geologic strata are introduced. The algorithms are extensions of the Smith and Waterman (1980) dynamic programming technique and include several features that greatly increase the utility for sedimentary sequences. Gaps in correlation (unconformities) caused by local nondeposition or eroded strata can include in a single "event" several strata. Furthermore, these gaps can be weighted as a single event, rather than as the sum of gap events for each strata. In addition, one or several adjacent strata in a second column can be correlated (matched) with one or several strata in a second column. Deletions within one of these multiple matches are also possible. The new algorithms include the method of minimum distance and the method of maximum similarity. Within this context, a similarity algorithm is given to locate and correlate the best matching segments or intervals from each stratigraphic column. All correlations within a preset distance of the optimum likewise can be produced for any of these algorithms. An example of specific assignments of these weight functions is given for correlation of well logs from the San Juan Basin.

KEY WORDS: stratigraphic correlation, well log correlation, dynamic programming, matching.

INTRODUCTION

The problem of correlating stratigraphic sequences is important in exploration and characterization of resources. These stratigraphic data can simply be sequences of lithologic units but may also include gamma logs, electric logs, strata thickness, geochemical or mineral arrays, or fossil occurrence and abundance. Until 1980, the most frequently used computer technique was cross-association, although other methods were attempted. [For cross-association see Sackin, Sneath, and Merriam (1965); Harbaugh and Merriam (1968); Merriam (1971); and Davis (1973). For other methods see Neidell (1969); Gill (1970); Matuszak (1972); Rudman and Lankston (1973); Shaw and Simms (1977); and

[1]Manuscript received 19 July 1985; accepted 10 July 1986.

[2]Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113.

[3]Earth & Space Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

0882-8121/87/0200-0109$05.00/1

Mann and Dowell (1978)]. Smith and Waterman (1980) presented a dynamic programming technique, the first one to overcome the problem of correlating across gaps. Gordon and Reyment (1979) and Gordon (1980) used dynamic programming but did not solve the gap problem.

At Los Alamos, we began work to extend the Smith and Waterman algorithm to include several features that would greatly increase the utility of the method. Howell, who assisted us, published a program written under our direction (Howell, 1983). Because our results as embodied in Howell's program were preliminary, we take this opportunity to report more complete results and to make some corrections. Important algorithms and modifications added to our early version are now discussed.

The first of these modifications utilizes multiple gaps. An essential property of these methods is the ability to include gaps in correlations. A single stratigraphic unit can be made a gap (not matched) and several adjacent units can be treated as a single gap (a multiple gap). The ability to include multiple gaps causes computation time to increase but this problem can be overcome in most cases (see the next section). Howell's (1983) code includes only single gaps.

A second modification to the early version is "many to many matching." Correlation of stratigraphic units can include several types of events. First, individual strata can be correlated or matched. Second, a stratigraphic unit from one column can be matched with several adjacent units in a second column. Howell's code includes these one-to-many matchings. One-to-many matching is related to the technique "time-warping" from speech recognition. [Kruskal and Liberman (1983) review this subject.] In addition, we introduce a new type of multiple matching, many-to-many matching, where a block of adjacent stratigraphic units in one column is matched to a block in a second column. Another new feature allows us to include gaps in the multiple matchings.

In addition to presenting minimum distance correlation (next section), we also introduce maximum similarity correlation (see section on maximum similarity correlation) and show how to convert from distance to similarity. A new technique to find most similar sections between two stratigraphic sequences is presented in the section on correlation of sections and is based on a modified similarity method.

Even the best correlation by a computer algorithm cannot always guarantee correct answers. For successful use of our technique, much care should be given to determining choice of weights for the algorithm. Still, even after weights are chosen carefully, the actual or desired correlation might only be near the "optimal," rather than be exactly optimal. To overcome this problem, we present a method (see section on finding optimal or near-optimal correlations) that produces all correlations within a user-specified distance of the computed "optimal."

In a section on algorithm specification, we discuss the rationale behind our choices for a specific example, some of which are concealed in R_{ij} or FUNCTION CUBE of Howell's code. Corrections to our earlier specifications as reported by Howell are also made in this section, and the extension to similarity is described. Different geological environments may out of necessity require different weighting schemes and one such example is discussed.

MINIMUM DISTANCE CORRELATION

To begin, let $\mathbf{a} = a_1 a_2, \ldots, a_n$ and $\mathbf{b} = b_1 b_2, \ldots, b_m$ be two stratigraphic sequences with n and m strata, respectively, where a_i or b_j represents lithologic characteristics of strata. To keep things general, we specify a gap function g and a distance function d. This keeps the initial formulas less cluttered and emphasizes the general applicability of the ideas. Associated with a gap of strata, $a_i a_{i+1}, \ldots, a_j$ ($j = i$ is possible) is a gap penalty, which we assume is positive

$$0 < g(a_i a_{i+1}, \ldots, a_j)$$

As mentioned above, we include matches or correlations of one or several stratigraphic units from one column with one or several from the other columns. One-to-one, one-to-many, many-to-one, and many-to-many can all be grouped together in one matching (distance) function d. Here $a_i a_{i+1}, \ldots, a_j$ is to be matched with $b_k b_{k+1}, \ldots, b_l$

$$0 \leq d(a_i a_{i+1}, \ldots, a_j, b_k b_{k+1}, \ldots, b_l)$$

This includes $j = i$ and $k = l$, $d(a_i, b_k)$, the case of matching single strata.

Next define the minimum distance $D(\mathbf{a}, \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$, the stratigraphic columns, to be the minimum weighted sum of matches and gaps required to correlate the columns. If d is a mathematical metric, then D can be shown to be also a metric and has the properties

(1) $D(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$

(2) $D(\mathbf{a}, \mathbf{b}) = D(\mathbf{b}, \mathbf{a})$ (symmetry)

(3) For any sequence $\mathbf{c}$, $D(\mathbf{a}, \mathbf{b}) \leq D(\mathbf{a}, \mathbf{c}) + D(\mathbf{c}, \mathbf{b})$ (triangle inequality)

The algorithm of Smith and Waterman (1980, their Eq. 5) covers only matching of single strata. The algorithm depends on considering

$$D_{ij} = D(a_1 a_2, \ldots, a_i, b_1 b_2, \ldots, b_j)$$

the correlation of $a_1 a_2, \ldots, a_j$ and $b_1 b_2, \ldots, b_j$. This correlation, if only one-to-one matching and single gaps are allowed, can end in one of three ways

a_1	b_1	a_1	b_1	a_1	b_1
a_2	b_2	a_2	b_2	a_2	b_2
⋮	⋮	⋮	⋮	⋮	⋮
a_{i-1}	b_j	a_i	b_{j-1}	a_{i-1}	b_{j-1}
a_i	$-\Delta$	Δ	$-b_j$	a_i	$-b_j$
(1) a_i gapped		(2) b_j gapped		(3) a_i and b_i matched	

Here, we focus on the bottom of the correlation. In case (1) where a_i is gapped, the earlier correlation has distance (by assumption) $D_{i-1,j}$. Therefore, case (1) has total distance $D_{i-1,j} + g(a_i)$.

In case (2), the distance is $D_{i,j-1} + g(b_j)$ whereas in case (3) the distance is $D_{i-1,j-1} + d(a_i, b_j)$. Hence the formula is

$$D_{ij} = \min \left[D_{i-1,j} + g(a_i),\ D_{i,j-1} + g(b_j),\ D_{i-1,j-1} + d(a_i, b_j)\right] \quad (1)$$

The recursion begins with $D_{00} = 0$, $D_{i0} = \Sigma_{l=1}^{i} g(a_i)$, $D_{0j} = \Sigma_{k=1}^{j} g(b_j)$.

Equation (1) extends to multiple gaps and multiple matching. For a gap of $a_l a_{l+1}, \ldots, a_i$, the corresponding term is

$$D_{l-1,j} + g(a_l a_{l+1}, \ldots, a_i), \qquad 1 \le l \le i$$

A gap of $b_k b_{k+1}, \ldots, b_j$ is handled in the same manner. For multiple matching of $a_l a_{l+1}, \ldots, a_i$ with $b_k b_{k+1}, \ldots, b_j$, the corresponding term is

$$D_{l-1,k-1} + d(a_l a_{l+1}, \ldots, a_i, b_k b_{k+1}, \ldots, b_j)$$

The extension of recursion Eq. (1) is

$$D_{ij} = \min \left\{ \begin{array}{l} \min\limits_{1 \le l \le i} \left[D_{l-1,j} + g(a_l, \ldots, a_i)\right]; \\ \min\limits_{1 \le k \le j} \left[D_{i,k-1} + g(b_k, \ldots, b_j)\right], \\ \min\limits_{\substack{1 \le l \le i \\ 1 \le k \le j}} \left[D_{l-1,k-1} + d(a_l, \ldots, a_i, b_k, \ldots, b_j)\right] \end{array} \right. \quad (2)$$

The algorithm is initialized by $D_{00} = 0$, $D_{i0} = g(a_1 a_2, \ldots, a_i)$ and $D_{0j} = g(b_1 b_2, \ldots, b_j)$.

The computer storage required for the D_{ij} array is nm, a reasonable requirement. The time to compute D_{nm} by Eq. (1) is $O(nm)$, a constant times nm. Unfortunately, this goes up to $O(n^2m^2)$ for the more useful Eq. (2). However, efficiency for Eq. (2) in most cases can be greatly improved to achieve nearly $O(nm)$ running time.

Before turning to algorithm efficiency, we add one more feature, deletions

within multiple matches. When matching $a_l, \ldots, a_i$ with $b_k, \ldots, b_j$, clearly most units $a_l, a_{l+1}, \ldots, a_i$ should be close to most of the units $b_k, b_{k+1}, \ldots, b_j$. Our distance function $d(a_l a_{l+1}, \ldots, a_i, b_k b_{k+1}, \ldots, b_j)$ should be severe regarding heterogeneity within $a_l, \ldots, a_i$ and $b_k, \ldots, b_j$. The following procedure is an approach which will accomplish our objective in reasonable computational time.

Suppose units $a_l, \ldots, a_i$ are classified into types $T_1, T_2, \ldots, T_N$. In the example shown later in the section on algorithm specification, 11 lithotypes are present. We also take a distance $\delta(T_s, T_t)$ on these types. Assign each a_x to type $T(a_x)$. Group all units close to type t into

$$S_t(E) = \left\{a_x : \delta[T(a_x), T_t] < E \quad \text{and} \quad l \le x \le i\right\}$$

where E is a preset constant. We must measure also the size of $S_t(E)$ and take the size to be in [0, 1]. The set $S_t = \varnothing$ should have size 0 whereas $S_t = \{a_l, a_{l+1}, \ldots, a_i\}$ should have size 1. In the example in the section on algorithm specification size is taken to be proportional to total thickness. Suppose $S_t(E)$ has the maximum size of $S_1(E), S_2(E), \ldots, S_N(E)$. If this maximum size is less than a preset value, say 0.9, deletions within multiple matches are not allowed for $a_l, a_{l+1}, \ldots, a_i$. If maximum size $S_t(E)$ is at least this preset value, then delete a_x not in $S_t(E)$. Of course, $b_k, \ldots, b_j$ is examined in an identical way.

Equation (2) above has two multiple gap terms, which contribute

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (i + j) = O(nm^2 + n^2 m)$$

to the computational cost of the algorithm. We now show how to reduce this to $O(nm)$ for a reasonable class of gap functions

$$g(a_l, \ldots, a_i) = \beta + \gamma \sum_{x=l}^{i} h(a_x)$$

where β and γ are nonnegative constants and h is a nonnegative function. That is, g is "linear" in $a_l, \ldots, a_i$. If we set

$$E_{i,j} = \min\left[D_{i,k-1} + g(b_k, \ldots, b_j) : 1 \le k \le j\right] \quad \text{and}$$

$$F_{ij} = \min\left[D_{l-1,j} + g(a_l, \ldots, a_i) : 1 \le l \le i\right]$$

then Eq. (2) becomes

$$D_{ij} = \min\left\{F_{ij}, E_{ij}, \min_{\substack{1 \le l \le i \\ 1 \le k \le j}} \left[D_{l-1,k-1} + d(a_l, \ldots, a_i, b_k, \ldots, b_j)\right]\right\}$$

Here $E_{00} = F_{00} = D_{00} = 0$, $E_{i0} = D_{i0} = g(a_1, \ldots, a_i)$, $F_{0j} = D_{0j} = g(b_1, \ldots, b_j)$. Next observe that

$$E_{ij} = \min\left\{D_{i,j-1} + g(b_j), \min\left[D_{i,k-1} + g(b_k, \ldots, b_j) : 1 \le k < j\right]\right\}$$
$$= \min\left\{D_{i,j-1} + g(b_j), \min\left[D_{i,k-1} + g(b_k, \ldots, b_j) : 1 \le k \le j-1\right]\right\}$$
$$= \min\left\{D_{i,j-1} + g(b_j), \min\left[D_{i,k-1} + g(b_k, \ldots, b_{j-1}) : 1 \le k \le j-1\right] + \gamma h(b_j)\right\}$$
$$= \min\left[D_{i,j-1} + g(b_j), E_{i,j-1} + \gamma h(b_j)\right]$$

Also

$$F_{ij} = \min\left[D_{i-1,j} + g(a_i), F_{i-1,j} + \gamma h(a_i)\right] \tag{4}$$

These two equations reduce gap calculations to $O(nm)$.

If the gap function is concave instead of linear, a somewhat more involved recursion also reduces calculation to $O(nm)$. Derivation for the concave gap function case can be obtained by modifying Waterman (1984). We assume the gap function has the form $g(a_l, \ldots, a_i) = f[\Sigma_{x=l}^{i} h(a_x)]$ and concave is, as usual, defined as $f(z+u) - f(z) \le f(y+u) - f(y)$, whenever $y \le z$ and $0 \le u$. We restrict study to $E_{i,j}$ because identical considerations hold for $F_{i,j}$. Now let

$$E_{i,j} = D_{i,l} + f\left[\sum_{x=l+1}^{j} h(a_x)\right] \le D_{i,k} + f\left[\sum_{x=k+1}^{j} h(a_x)\right]$$

for all $0 \le k \le j-1$. For $l \le k$

$$f\left[\sum_{x=l+1}^{j+1} h(a_x)\right] - f\left[\sum_{x=l+1}^{j} h(a_x)\right]$$
$$\le f\left[\sum_{x=k+1}^{j+1} h(a_x)\right] - f\left[\sum_{x=k+1}^{j} h(a_x)\right]$$

and, adding the last two inequalities

$$D_{i,l} + f\left[\sum_{x=l+1}^{j+1} h(a_x)\right] \le D_{i,k} + f\left[\sum_{x=k+1}^{j} (a_x)\right]$$

The implications are that

$$E_{i,j+1} = \min\left[D_{i,k} + g(a_{k+1}, \ldots, a_{j+1}) : 0 \le k \le l;\ D_{i,j} + g(a_{j+1})\right]$$

Therefore, concave gap functions can reduce computation. Moreover, the above implies that minimization can be related to the set

$$S(i) = \{k : E_{i,k+1} = D_{i,k} + g(a_{k+1})\} \quad \text{and}$$

$$E_{i,j} = \min \left[D_{i,k} + g(a_{k+1}, \ldots, a_j) : k \in S(i)\right]$$

This does not obtain the $O(nm)$ result of linear gap functions, but that is the practical implication.

We define a gap function to be convex if $f(y + u) - f(y) \leq f(z + u) - f(z)$, whenever $y \leq z$ and $0 \leq u$. Argument similar to the last paragraph shows

$$E_{i,j+1} = \min \left[D_{i,k} + g(a_{k+1}, \ldots, a_{j+1}) : l \leq k \leq j\right]$$

again allowing reduction in the computation. In fact, this equation immediately gives an $O(nm)$ algorithm.

The major contribution to computation time in Eq. (2) is not from gap terms but from the multiple matching term

$$\min_{\substack{1 \leq l \leq i \\ 1 \leq k \leq j}} \left[D_{l-1,k-1} + d(a_l, \ldots, a_i, b_k, \ldots, b_j)\right]$$

This term takes computation time proportional to

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (i \cdot j) = 0(n^2m^2)$$

This was reduced in our earlier algorithm as reported by Howell (see the section on algorithm specification) by the simple device of only multiple matching $a_l, \ldots, a_i$ and $b_k, \ldots, b_j$ if the total thickness of $a_l, \ldots, a_i$ is close ($\pm 50\%$) to that of $b_k, \ldots, b_j$. This essentially reduces $O(n^2m^2)$ to $O[\min (n^2m, nm^2)]$.

MAXIMUM SIMILARITY CORRELATION

We shall retain our notation for two stratigraphic sequences $\mathbf{a} = a_1a_2, \ldots, a_n$ and $\mathbf{b} = b_1b_2, \ldots, b_m$. We derived algorithms for minimum distance correlation, basing the approach on finding the correlation with the least difference between **a** and **b** see previous section. Here we use the equally intuitive idea of finding the correlation with the most similarity between **a** and **b**. After presenting the basic algorithm, we show how to build similarity functions from distance functions (and vice versa) which yield equivalent optimal correlations. These two approaches might seem to be different ways of saying the same thing, replacing minimum by maximum. However, the issue of which approach to use is not simply a matter of taste; we present in the section on correlation of sections a similarity based algorithm for a problem which has no equivalent distance algorithm.

To distinguish a new gap function of this section, we use g^*

$$0 < g^*(a_ia_{i+1}, \ldots, a_j)$$

The similarity functions between strata is not assumed strictly positive or strictly negative; frequently the similarity function s is taken to be both. Let

$$s(a_i a_{i+1}, \ldots, a_j, b_k b_{k+1}, \ldots, b_l)$$

be given for $1 \le i \le j \le n$ and $1 \le k \le l \le n$. In this setting, we wish to accumulate similarity, so s should be positive for good matches and negative for poor ones. The similarity function S maximizes, over all correlations C, the sum of similarity values of matches minus gap penalties

$$S(a, b) = \max_C \left[\sum_{\substack{a \& b \\ \text{matched}}} s(a, b) - \sum_{\substack{e \\ \text{gapped}}} g^*(e) \right]$$

where C ranges over all correlations. If $S_{ij} = S(a_1, \ldots, a_i, b_1, \ldots, b_j)$, the single gap, single match algorithm (1) becomes

$$S_{ij} = \max \left[S_{i-1,j} - g^*(a_i), S_{i,j-1} - g^*(b_j), S_{i-1,j-1} + s(a_i, b_j) \right] \quad (5)$$

The multiple match, multiple gap algorithm is

$$S_{ij} = \max \left\{ \begin{array}{l} \max\limits_{1 \le l \le i} \left[D_{l-1,j} - g^*(a_l, \ldots, a_i) \right]; \\ \max\limits_{1 \le k \le j} \left[D_{i,k-1} - g(b_k, \ldots, b_j) \right]; \\ \max\limits_{\substack{1 \le l \le i \\ 1 \le k \le j}} \left[D_{l-1,k-1} + s(a_l, \ldots, a_i, b_k, \ldots, b_j) \right] \end{array} \right\} \quad (6)$$

The efficiencies derived earlier carry over to these algorithms as well.

Distance correlation and similarity correlation are equivalent. We conclude this section by demonstrating this for the single match, single gap case. Although true in general, the single gap, single match requires much less notation. In any correlation C, $n + m$ units occur from the two columns. The equation

$$n + m = 2 \text{ (number matches)} + \text{number gaps}$$

is obviously true. From the distance function $d(a, b)$, construct a similarity function

$$s(a, b) = K - d(a, b)$$

where K is a fixed constant, $0 \le K \le \max_{a,b} d(a, b)$. Then

$$D(a, b) = \min_C \left[\sum_{\text{matches}} d(a, b) + \sum_{\text{gaps } e} g(e) \right]$$

$$= \min_C \left[\sum_{\text{matches}} K - \sum_{\text{matches}} s(a, b) + \sum_{\text{gaps}\, e} g(e) \right]$$

$$= \min_C \left[K \frac{n+m}{2} - \frac{K}{2} \text{number gaps} - \sum_{\text{matches}} s(a, b) + \sum_{\text{gaps}\, e} g(e) \right]$$

$$= K \frac{n+m}{2} - \max_C \left\{ \sum_{\text{matches}} s(a, b) - \sum_{\text{gaps}\, e} [g(e) - K/2] \right\}$$

This says that if

$$s(a, b) = K - d(a, b)$$

then

$$g^*(e) = g(e) - K/2$$

and, moreover

$$D(\mathbf{a}, \mathbf{b}) + S(\mathbf{a}, \mathbf{b}) = K[(n+m)/2]$$

Thus for any constant K, the above equations show that for any distance (similarity) correlation problem, a similarity (distance) correlation problem exists with identical optimal correlations.

CORRELATION OF SECTIONS

Frequently in geologic strata, a surprisingly large section or segment of one column will correlate well with a section from the second column, whereas the remainder of the strata correlate poorly. As a result, methods presented above (minimum distance and maximum similarity) may fail to find good quality correlation. In an earlier paper, Smith and Waterman (1980) show how to modify the minimum distance algorithm to correlate a fragmentary sequence with a longer complete sequence. However, that idea can not be extended to solve the more general problem described here.

The difficulty with extending the distance algorithm to a best sections algorithm is shown in the following simple example. Both $D(a, a) = 0$ and $D(aaa, aaa) = 0$ where a represents some lithologic unit. Therefore, distance does not distinguish short identities from long identities. However, similarity gives larger scores for longer matches: $S(a, a) = s(a, a)$ whereas $S(aaa, aaa) = 3s(a, a)$. This property was exploited in an algorithm of Smith and Waterman (1981a, b) and is adapted for use in the present context.

Values of $s(a, b)$ are assumed to be set such that negative values correspond to poor quality matches and positive values to good quality matches. Define

$$H_{ij} = \max\left[0,\ S(a_x a_{x+1}, \ldots, a_i, b_y b_{y+1}, \ldots, b_j):\right.$$
$$\left.1 \le x \le i \quad \text{and} \quad 1 \le y \le j\right]$$

Fortunately, a simple algorithm to find H_{ij} exists. For the single gap, single match case

$$H_{ij} = \max\left[0,\ H_{i-1,j} - g^*(a_i),\ H_{i,j-1} - g^*(b_j),\ H_{i-1,j-1} + s(a_i, b_j)\right]$$
$$\text{where } H_{i0} = H_{0j} = 0 \quad \text{for} \quad 0 \le i \le n \quad 0 \le j \le m \tag{7}$$

This algorithm, as usual, is easily extended to the multiple gap, multiple match case

$$H_{ij} = \max\left\{\begin{array}{l} 0;\ \max\limits_{1 \le l \le i}\left[D_{l-1,j} - g^*(a_l, \ldots, a_i)\right] \\ \max\limits_{1 \le k \le j}\left[D_{i,k-1} - g^*(b_k, \ldots, b_j)\right] \\ \max\limits_{\substack{1 \le l \le i \\ 1 \le k \le j}}\left[D_{l-1,k-1} + s(a_l, \ldots, a_i, b_k, \ldots, b_j)\right] \end{array}\right\} \tag{8}$$

where $H_{0j} = H_{i0} = 0$ for $0 \le i \le n$ and $0 \le j \le m$. Again, efficiencies (see section on minimum distance correlations) carry over to this situation.

FINDING ALL OPTIMAL AND NEAR-OPTIMAL CORRELATIONS

All methods presented above are designed to find the distance and or similarity corresponding to an optimal correlation. However, we have not discussed how the actual correlations, which are of primary interest, can be produced. Although progamming for finding correlations is slightly more involved than for the forward recursion, computation time is small compared to the forward recursion.

Initially, we show how to find the set of all optimal correlations. Then we adapt some recent results in dynamic programming to show how to produce *all* correlations within a user-specified distance of the optimum. Even when much care is given to establishing weights for the algorithm, the correct correlation is not always optimal. If, however, the correlation is close to optimal, our near-optimal algorithm will produce it.

Although all optimal correlations can be produced by setting pointers on the forward pass, we prefer a more straightforward approach, which we refer to as a traceback. For definiteness, assume we are doing a single gap, single match distance correlation. We have calculated the matrix

$$(D_{ij}) \begin{array}{l} j = 1, \ldots, m \\ i = 1, \ldots, n \end{array}$$

Optimal correlations correspond to paths through this matrix, which we begin at D_{nm}. Assume we have discovered an optimal correlation leading from (n, m) to, but not including (i, j). We can discover easily what step(s) led us to compute D_{ij} on the forward recursion by asking three questions

$$\begin{aligned}
&(1)\ D_{i-1,j} + g(a_i) = D_{ij}?\\
&(2)\ D_{i,j-1} + g(b_j) = D_{ij}?\\
&(3)\ D_{i-1,j-1} + d(a_i, b_j) = D_{ij}?
\end{aligned} \tag{9}$$

If, for example, $D_{i-1,j} + g(a_i) = D_{ij}$, we know a_i can be gapped and we repeat the three questions using $(i - 1, j)$. Multiple optimal correlations can be obtained by remembering (stacking) unexplored directions.

Recent work (Waterman, 1983 and more fully described in Byers and Waterman, 1985) allows us to find all correlations within e of the optimum. This time we assume that at position (i, j) a correlation is being generated that can result in a total alignment sum less than or equal to $D_{nm} + e$. The alignment sum from (n, m) to but not including (i, j) is T_{ij}. The three questions above become

$$\begin{aligned}
&(1)\ D_{i-1,j} + g(a_i) + T_{ij} \leq D_{nn} + e?\\
&(2)\ D_{i,j-1} + g(b_j) + T_{ij} \leq D_{nm} + e?\\
&(3)\ D_{i-1,j-1} + d(a_i, b_j) + T_{ij} \leq D_{nm} + e?
\end{aligned} \tag{10}$$

To illustrate these ideas we take an elementary textbook example which was also studied in Smith and Waterman (1980). There

$$d(a, b) = \begin{cases} 0 & a = b \\ 3 & a \neq b \end{cases}$$

whereas $g(a) = 1$ for all a. The sequences from Table 1 have $D(a,b) = 5$. The single optimal correlation, including the five gaps which results in a total distance of 5, is determined (Table 1)

E	*E*
F	*F*
Δ	*A*
B	*B*
F	*F*
E	*E*
F	*F*
B	*B*
E	Δ
F	Δ

Table 1. Distance Correlation Between Two Outcrops[1]

	0	*E*	*F*	*A*	*B*	*F*	*E*	*F*	*B*	*A*	*B*	*F*	*E*	*F*	*B*
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
E	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
F	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12
B	3	2	1	2	1	2	3	4	5	6	7	8	9	10	11
F	4	3	2	3	2	1	2	3	4	5	6	7	8	9	10
E	5	4	3	4	3	2	1	2	3	4	5	6	7	8	9
F	6	5	4	5	4	3	2	1	2	3	4	5	6	7	8
B	7	6	5	6	5	4	3	2	1	2	3	4	5	6	7
E	8	7	6	7	6	5	4	3	2	3	4	5	4	5	6
F	9	8	7	8	7	6	5	4	3	3	5	4	5	4	5
A	10	9	8	7	8	7	6	5	4	3	4	5	6	5	6
B	11	10	9	8	8	8	7	6	5	4	3	4	5	6	5
F	12	11	10	9	9	8	8	7	6	5	4	3	4	5	6
E	13	12	11	10	10	9	8	8	7	6	5	4	3	4	5

[1]Distance correlation between a stratigraphic sequence *EFABFEFBABFEFB* in a Wynandotte County, Kansas roadcut with sequence *EFBFEFBEFABFE* in an Osage County, Kansas roadcut; *F* = limestone, *E* = marine shale, *D* = coal, *C* = underclay, *B* = nonmarine shale, and *A* = sandstone. Data from Harbaugh and Merriam, 1968, p. 173.

A	A
B	B
F	F
E	E
Δ	F
Δ	B

ALGORITHM SPECIFICATION: AN EXAMPLE

The success of algorithms presented above is dependent heavily on choices of weight functions, $g(a_i, \ldots, a_j)$, $d(a_i, \ldots, a_j, b_k, \ldots, b_l)$, and $s(a_i, \ldots, a_j, b_k, \ldots, b_l)$. Functions which are satisfactory for one application will prove to be inadequate for others. Within this section, we present some of the reasons for functions we chose in correlating sequences from the San Juan Basin. The actual functions used in Howell (1983) are described, corrected, and extended to similarity correlation.

We obtained the San Juan data from a report of Beach and Jentgen (1978), a reference missing in Howell (1983). The stratigraphic sequences were converted into the form

$$a_i = [\text{Major lithologic type; minor lithologic type (modifying lithologic adjective); strata thickness}] = (t_i^1, t_i^2, w_i)$$

The major lithologic types encountered were coal (COA), limestone (LIM), sandstone (coarse-, SSC; medium-, SSM; fine-, SSF, and very fine-, SSV grained), siltstone (SLT), shale (SHA), claystone (CLA), and alluvium (ALL). All of the above could likewise be used as minor lithologic types for modifying major types. Carbonaceous (CAR) was used only as a minor lithologic type or modifier. Any combination of major and minor lithotypes were possible. For example, clayey medium-grained sandstone = (SSM, CLA); coaly shale = (SHA, COA); or carbonaceous fine-grained sandstone = (SSF, CAR). These three letter abbreviations are used below.

Distance between two stratigraphic units a_i and b_j will be considered first, before multiple matching. Clearly, type (t_i) differences and thickness (w_i) differences must be blended into a single distance measure. Let $a_i = (t_i^1, t_i^2, w_i)$ and $b_j = (u_j^1, u_j^2, v_j)$. Consider first the distance between two lithologic types t and u. Sandstone, siltstone, and claystone lithologic types are defined by ϕ values where $\phi = \log_2(d)$ and d is the diameter of grain size. For these lithologic types, we define distance $\delta(t, u)$ between t and u to be

$$\delta(t, u) = |\phi(t) - \phi(u)|$$

Coal, carbonaceous (CAR), and limestone are defined by chemical com-

position rather than by ϕ size. Alluvium similarly is not defined by ϕ size, but rather, by origin—relating to the fact that it is recent, unconsolidated material overlying older rocks. Except for coal relative to carbonaceous (which are both rich in organics), we have chosen all four of these lithotypes to have the maximum defined distance, $\max \delta(t, u) = 6$, from all other lithotypes, due to great dissimilarity based on factors other than ϕ size (Table 2). Shale (SHA), which by definition is composed of grains with a ϕ size equivalent to claystone, is defined to be the same distance as claystone from all other lithotypes. However, as a result of textural differences, shale is more laminated and fissle than claystone, and a minimum distance, $\delta(t, u) = 0.5$, between shale and claystone is used.

The next task is to combine δ values for major and minor types into one distance measure

$$\delta^*(t_i^1 t_i^2; u_j^1 u_j^2) = \alpha\delta(t_i^1, u_j^1) + (1 - \alpha)\,\delta(t_i^2, u_j^2) \tag{11}$$

Here we have simply taken a linear combination of the major and minor type $\delta(t, u)$ distances [$\alpha = q_1^*$, eq. (5) of Howell, 1983].

Only strata thickness remains to be added to our distance function. Our earlier measure was

$$|w_i - v_j|/q_2$$

where, for example, $q_2 = 10$ ft, say. This would put every 10 ft of difference between strata thickness into equal weight with one unit of ϕ distance between lithotypes. What is troubling about this assignment is that $|10 - 20|/10 = 1$ is equivalent to $|110 - 120|/10 = 1$, so that absolute difference is the consideration. A much more reasonable assignment seems to be

$$\frac{|w_i - v_j|}{\max\{w_i, v_j\}} \cdot \max\{\delta(t, u)\} \tag{12}$$

This gives thickness distance a value which is relative to individual strata thickness and in the same range as lithotype distance.

We combine Eqs. (11) and (12) into

$$\begin{aligned} d(a_i, b_j) &= d[(t_i^1, t_i^2, w_i), (u_j^1, u_j^2, v_j)] \\ &= \alpha\delta(t_i^1, u_j^1) + (1 - \alpha)\,\delta(t_i^2, v_j^2) \\ &\qquad + \frac{|w_i - v_j|}{\max\{w_i, v_j\}} \max_{t,u} \delta(t, u) \\ &= \alpha\delta(t_i^1, u_j^1) + (1 - \alpha)\,\delta(t_i^2, v_j^2) + 6\frac{|w_i - v_j|}{\max\{w_i, v_j\}} \end{aligned} \tag{13}$$

Table 2. Distance Function $\delta(t, u)$ Between Lithotypes t and u

	COA	LIM	SSC	SSM	SSF	SSV	SLT	CLA	SHA	CAR	ALL
COA	0	6	6	6	6	6	6	6	6	4	6
LIM	6	0	6	6	6	6	6	6	6	6	6
SSC	6	6	0	1	2	3	4	5	5	6	6
SSM	6	6	1	0	1	2	3	4	4	6	6
SSF	6	6	2	1	0	1	2	3	3	6	6
SSV	6	6	3	2	1	0	1	2	2	6	6
SLT	6	6	4	3	2	1	0	1	1	6	6
CLA	6	6	5	4	3	2	1	0	0.5	6	6
SHA	6	6	5	4	3	2	1	0.5	0	6	6
CAR	4	6	6	6	6	6	6	6	6	0	6
ALL	6	6	6	6	6	6	6	6	6	6	0

Equation (13) defines one-to-one matching. It must be extended to multiple matches (one-to-many and many-to-many). The idea behind multiple matching is to blend several similar strata into one somewhat heterogeneous stratum. Suppose individual units are $a_i a_{i+1}, \ldots, a_j$ and $b_k b_{k+1}, \ldots, b_l$. We extend Eq. (11) by

$$\delta^*(a_i a_{i+1}, \ldots, a_j; b_k b_{k+1}, \ldots, b_l) = \frac{\sum_{x=i}^{j} \sum_{y=k}^{l} \delta(a_x, b_y)}{\min\{j - i + 1, l - k + 1\}} \tag{14}$$

Dividing by the minimum number of units in either $a_i, \ldots, a_j$ or $b_k, \ldots, b_l$, lets us view this correlation of the larger number of units with the "average" unit in the opposite column. Notice that this coincides precisely with the one-to-one correlation distance in Eq. (11) above or the one-to-many equation presented in our earlier work. We also suggest considering weighting $\delta(a_x, b_y)$ by the proportion of type $a_x(b_y)$ in the respective columns.

The extension, then, of Eq. (13) to general multiple matching, is

$$d(a_i, \ldots, a_j, b_k, \ldots, b_l) = \frac{\sum_{x=i}^{j} \sum_{y=k}^{l} [\alpha\delta(t_x^1, u_y^1) + (1 - \alpha)\,\delta(t_x^2, u_y^2)]}{\min\{j - i + 1, l - k + 1\}} + \frac{6\left|\sum_{x=i}^{j} w_x - \sum_{y=k}^{l} v_y\right|}{\max\left\{\sum_{x=i}^{j} w_x, \sum_{y=k}^{l} v_y\right\}} \tag{15}$$

Gaps have not been discussed yet. Reasonably, an initial gap cost β and a per unit of thickness cost γ may be defined. This gives a gap function of the form

$$g(a_i) = \beta + \gamma w_i$$

which readily extends to

$$g(a_i a_{i+1}, \ldots, a_j) = \beta + \gamma \sum_{x=i}^{j} w_x$$

This function is "linear" and allows the efficiencies discussed earlier.

The final issue we wish to address is conversion of these distance functions to those appropriate for similarity. The key is the choice of the constant K (see section on maximum similarity correlation) where

$$s(a, b) = K - d(a, b)$$

We would like $s(a, b) > 0$ for "quality" correlations, $s(a, b) = 0$ for correlations we are indifferent about, and $s(a, b) < 0$ for poor correlations. K is, obviously, a parameter the user can adjust. To achieve both positive and negative values of s, K must satisfy

$$0 < K < \max_{a,b} d(a, b)$$

For single strata a and b

$$\max_{a,b} d(a, b) = \max_{t,u} \delta(t, u) + \max_{w,v} \frac{6|w - v|}{\max\{w, v\}}$$

$$= 6 + 6 = 12$$

Any value of K larger than 6 does not seem useful to us, whereas a value of K as small as 3 seems fairly restrictive.

SUMMARY

A need for more rapid quantitative correlation of sedimentary strata, whether dealing with outcrops, cores, or geophysical log data has existed. However, with realization that correlation of geologic sections with no gaps is the exception rather than the rule, an approach had to be developed that would consider all anomalies present between two correlative geologic sections. Such an approach had to account for erosional disconformities, emplacement in one section and not in the other of isolated or convulsive depositional events, and occurrence of sedimentary facies representing contemporaneous deposition of more than one sediment type in adjoining environments. The approach described in this paper accounts for such anomalies, respectively, though the ability to include single and multiple gaps in the correlations, through the ability to match a stratigraphic unit from one section with several adjacents units in the second section, and through the ability to compare both minimum distance and maximum similarity within a single correlation. Furthermore, based on one's knowledge of various depositional environments, different weighting schemes may be applied to account for chemical and physical variations occurring in the strata under study.

We do not suggest that the proposed algorithms for stratigraphic correlation replace the geologist or geophysicist presently correlating sedimentary sections. Indeed, the geologist/geophysicist must determine the weighting schemes for correlations and, based on those schemes, make a decision as to the most reasonable of the multiple best fits that result. We have shown that algorithms

exist, that when combined with the knowledge of the investigator, provide a mechanism to consider complex sedimentary phenomena in automated correlations.

ACKNOWLEDGMENTS

The authors are apologetic that they have so long delayed reporting these results and are grateful to Jo Ann Howell for publishing some of our preliminary results. As reported in her paper, she translated an earlier version of some of these algorithms into FORTRAN. Temple Smith assisted her with the "one-to-many" summation and we appreciate his help in obtaining usable code. The early work was supported by a New Research Initiatives Award from Los Alamos National Laboratory to Michael Waterman and Robert Raymond. Later development reported in this paper was supported by Schlumberger and the System Development Foundation.

REFERENCES

Agterberg, F. P. and Gradstein, F. M., 1981. Workshop on Quantitative Stratigraphic Correlation Techniques: Math. Geol., v. 13, p. 81–91.

Beach, L. J. and Jentgen, R. W., 1978. Coal test drilling for the San Juan Mine Extension, San Juan County, New Mexico, U.S. Department of the Interior Geological Survey Open-File Report 78-960.

Byers, T. H. and Waterman, M. S., 1984. Determining all optimal and near-optimal solutions when solving shortest path problems by dynamic programming: Oper. Res., v. 32, p. 1381–1384.

Davis, John C., 1973. Statistics and Data Analysis in Geology: John Wiley & Sons, New York, p. 550.

Gill, Dan, 1970. Application of a Statistical Zonation Method to Reservoir Evaluation and Digitized Log Analysis: Amer. Assoc. Pet. Geol. Bull., v. 54, n. 5, p. 719–729.

Gordon, A. D., 1980. A FORTRAN IV Program for Comparing Two Sequences of Observations: Comput. Geosci., v. 6, n. 1, p. 7–20.

Gordon, A. D. and Reyment, R. A., 1979. Slotting of Borehole Sequences: Math. Geol., v. 11, p. 309–327.

Harbaugh, John W. and Merriam, Daniel F., 1968. Computer Applications in Stratigraphic Analysis: John Wiley & Sons, New York, 282 p.

Howell, J. A., 1983. A FORTRAN 77 Program for Automatic Stratigraphic Correlation: Comput. Geosci., v. 9, p. 311–327.

Kemp, Franklin, 1982. An Algorithm for the Stratigraphic Correlation of Well Logs: Math. Geol., v. 14, p. 271–285.

Kruskal, J. B. and Liberman, N., 1983. The Symmetric Time-warping Problem: From Continuous to Discrete, in Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, *in* Sankoff, D. and Kruskal, J. B. (eds.): Addison Wesley, London, p. 125–161.

Mann, C. John and Dowell, Thomas P. L., 1978. Quantitative Lithostratigraphic Correlation of Subsurface Sequences: Comput. Geosci., v. 4, n. 3, p. 295–306.

Matuszak, D. R., 1972. Stratigraphic Correlation of Subsurface Geological Data by Computer: Math. Geol., v. 4, p. 331–343.

Merriam, Daniel F., 1971. Computer applications in stratigraphic problem solving in decision making in the mineral industry: Can. Inst. Min. Met., v. 12, p. 139-147.

Neidell, N. S., 1969. Ambiguity Functions and the Concept of Geological Correlation: Kansas Geol. Survey Comput. Contr., v. 40, p. 19-29.

Preston, Floyd W. and Henderson, James H., 1964. Fourier Series Characterizations of Cyclic Sediments for Stratigraphic Correlation: Kansas Geol. Surv. Bull. 169, v. II, p. 415-425.

Rudman, A. J. and Lankston, R. W., 1973. Stratigraphic Correlation of Well Logs by Computer Techniques: Amer. Assoc. Pet. Geol. Bull., v. 57, n. 3, p. 577-588.

Sackin, M. J., Sneath, P. H. A., and Merriam, Daniel F., 1965. ALGOL Program for Cross-association of Non-numeric Sequences Using a Medium Size Computer: Kansas Geol. Surv. Sp. Dist. Publ., v. 23, 36 p.

Schwarzacher, Walter, 1980. Models for the Study of Stratigraphic Correlation: Math. Geol., v. 12, p. 213-234.

Shaw, Brian R. and Simms, R., 1977. Stratigraphic Analysis System: SAS: Comput. Geosci., v. 3, n. 3, p. 395-427.

Smith, T. F. and Waterman, M. S., 1980. New Stratigraphic Correlation Techniques: Jour. Geol., v. 88, n. 4, p. 451-457.

Smith, T. F. and Waterman, M. S., 1981a. Identification of Common Molecular Sequences: J. Mol. Biol., v. 147, p. 195-197.

Smith, T. F. and Waterman, M. S., 1981b. Comparison of Biosequences: Adv. Appl. Math., v. 2, p. 482-489.

Waterman, M. S., 1983. Sequence Alignments in the Neighborhood of the Optimum with General Applications to Dynamic Programming: Proc. Natl. Acad. Sci. U.S.A., v. 80, p. 3123-3124.

Waterman, M. S., 1984. Efficient Sequence Alignment Algorithms: J. Theor. Biol., v. 108, p. 333-337.