Mathematical Geology, Vol. 38, No. 7, November 2006 (© 2007)
DOI: 10.1007/s11004-006-9062-8

Graphic Biostratigraphic Correlation Using Genetic Algorithms[1]

Tao Zhang[2] and Roy E. Plotnick[3]

The most generally used method for estimating the basin-wide sequence and scaling of first and last occurrences, based on their occurrence in local sections, is Shaw's graphic correlation method. The key step in this method is the determination of the line of correlation (LOC), which represents the best estimate of the correlation between two local sections, or between a local section and a composite standard. In general, available techniques for fitting the LOC for multiple sections are tedious, subjective, or computationally expensive. A new method employing genetic algorithms can dramatically reduce the effort involved in determining the LOC and produces stable biostratigraphic correlations and composite range charts objectively and efficiently. Genetic algorithms are an artificial intelligence technique that excels in locating the optimum solution from a large number of alternative choices. In the case of the LOC, the alternative choices are the number of line segments comprising the complete line and the positions of each segment's beginning and end points. For a given number of segments, a wide range of alternative LOCs can be rapidly evaluated and a potential optimum fit determined. It is also possible to estimate the point when no further refinement of the fit by adding line segments is necessary. Genetic algorithms can also be applied to other methods for quantitative biostratigraphy.

KEY WORDS: biostratigraphy; correlation; artificial intelligence; genetic algorithms.

INTRODUCTION

Quantitative biostratigraphy uses the measured fossil record in sedimentary rock sections for the reconstruction of biological events in time and space (Sadler, 2004). It aims to organize local records of fossil occurrence and produce the best possible correlations among them. Several approaches to numerical correlation methods have been developed; they attempt to produce a consistent temporal framework for stratigraphic analysis of a basin (Agterberg, 1990). The development of these methods has been promoted by the widespread availability of high speed computers.

[1]Received 24 February 2005; accepted 26 October 2005; Published online: 14 February 2007.
[2]Wireless Broadband Systems, Motorola, Inc., 1475 W Shure Drive, Arlingtong Heights, IL 60004, USA; e-mail: CTZ020@motorola.com
[3]Department of Earth and Environmental Sciences, University of Illinois at Chicago, 845 W. Taylor St. Chicago, IL 60607, USA; e-mail: plotnick@uic.edu

A parallel development is the application of artificial intelligence (AI) methods to solve complex problems (Patterson, 1990; Winston, 1992). AI attempts to mimic human thinking and reasoning processes. An artificial intelligence model can be trained or self-taught to memorize a set of given rules and to generalize the rules to evaluate and respond to new input information. Artificial intelligence methods are employed to solve non-linear problems that are not suitable for the application of analytic or conventional statistical techniques.

This paper describes how genetic algorithms can be integrated with the graphic correlation method of Shaw (1964) to produce an optimized fit of the line-of-correlation for comparisons among stratigraphic sections and the production of composite standard sections. The method will be evaluated with available biostratigraphic data sets and validated using a basin simulation model.

Graphic Correlation

One of the primary goals of quantitative biostratigraphy is to use the order of biostratigraphic events seen in local sections to reconstruct their actual original sequence (Agterberg, 1990). Toward this goal, efforts have been made to develop automated stratigraphic correlation tools (Agterberg, 1982; Nel, 1982a,b; Edwards, 1978, 1984, 1995; Grimm, 1987; Kemple and others, 1995; Sadler and Cooper, 2003; Sadler, 2004). Many of these tools share the underlying logic of the graphic correlation method of Shaw (1964, 1995; Miller, 1977; Mann and Lane, 1995).

Graphic correlation estimates the global sequence and spacing of first appearance datums (FAD's or bases) and last occurrence datums (LAD's or tops), based on their occurrence in local sections. Graphic correlation solves ranking and scaling of biostratigraphic occurrences by including assumptions about the relationship between stratigraphic thickness and time. The technique thus can be used to estimate true total stratigraphic ranges from local stratigraphic ranges (Carney and Pierce, 1995). These ranges, in turn, can be the basis of global diversity estimates (Sadler and Cooper, 2003).

The basic method for graphic correlation is simple and intuitive (Macleod and Sadler, 1995; Carney and Pierce, 1995). A two-dimensional graph is constructed with one section, usually one with the most complete and detailed sampling and is unfaulted, the *standard reference section* (SRS), is placed on the *X*-axis (Fig. 1). A second section is placed along the *Y*-axis. The bases of the two sections are at the origin. The coordinates of fossil last and first occurrences (tops and bases) are now plotted in the plane of the graph. These are potential points of correlation between the two sections; based on various criteria, the most reliable of these are chosen as points of correlation (Carney and Pierce, 1995). Based on these points of correlation, a *line of correlation* (LOC) is drawn between the two sections. The LOC represents an estimate of the true correlation that exists between the

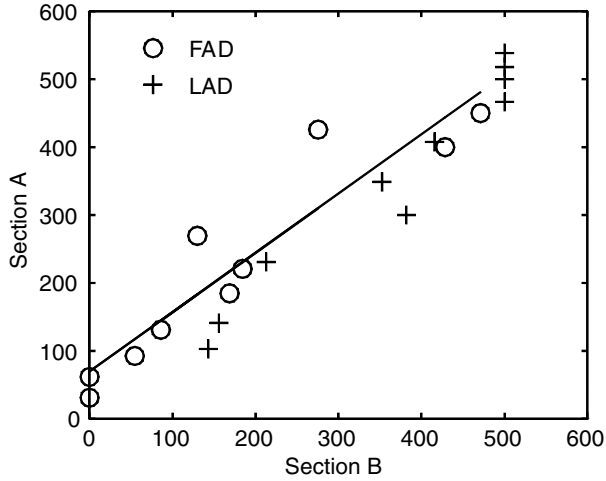

Figure 1. Example data set, from Miller (1977), with the line of correlation, fit by a simple linear regression of locations in Section A on those in Section B.

sections. The LOC shown in Figure 1 is a simple linear regression fit to the tops and bases.

Once determined, the LOC is used to project the tops and bottoms of the second section to the reference section. The reference section is now modified so that ranges are extended downward if the projected bottom is lower than that of the original bottom and extended upward if the projected top is higher. The result is a composite standard (CS) reference section, containing information from both original sections. The projections from additional local sections can be combined with the CS in the same way, producing further refinement of the CS.

The technique is usually carried out in multiple rounds (Carney and Pierce, 1995). After the CS has been produced in the first round, a second round of graphing plots the individual sections against the CS, with the data from that section removed beforehand. The final CS is the best overall estimate of the sequence of biostratigraphic events based on all available sections.

Fitting the LOC

As pointed out by many authors (e.g., Edwards, 1995; MacLeod and Sadler, 1995) the critical step in graphic correlation is the placement of the LOC. For two time correlative sections, a true LOC or "snake" of correlation must exist (Edwards, 1995; Kemple and others, 1995). However, this true line of correlation

is, by nature, unknowable (akin to a population in statistics) and can only be estimated (Edwards, 1984).

How to draw this estimated LOC as objectively as possible, without violating geological rules (such as superposition) and knowledge, is the leading problem for graphic correlation. The chief issues are how many line segments (doglegs) should comprise the estimated LOC, and how these lines should be placed. A major constraint is that the line cannot have a negative slope.

Following previous work, especially Edwards (1984, 1995), the desiderata for an ideal LOC fitting procedure can be summarized as follows:

(1) Although some subjectivity is unavoidable, it should be as minimal and as clearly laid out as possible.
(2) The geometry of the LOC must meet all relevant geological observations and concepts; i.e, it should not violate the concept of superposition.
(3) The procedure should be readily reproducible.
(4) The procedure should be able to keep computing time to a minimum and efficiently handle multiple sections.
(5) Reliability evaluations (weighting) of individual datums (Edwards, 1995) should be included.
(6) Include as many line segments as are needed, but no more than are needed (Edwards, 1995; MacLeod and Sadler, 1995)
(7) The LOC fitting method should be stable and robust. Small changes or additions in data should not significantly change the overall results; at the same time, substantial new and high quality data should have the potential to significantly change the LOC.

Quantitative approaches to fitting the LOC, including Shaw's original method (1964) and its revised versions by Miller (1977) and Edwards (1984, 1995), are usually based on least squares regression, although other methods of line fitting, such as reduced major axis are available (Macleod and Sadler, 1995). The simplest situation is when a single straight line is fit to the data. Figure 1 shows a linear regression, for the example data set of Miller (1977), of Section A on Section B. A single line of correlation is only valid, however, when the ratios of sediment-accumulation rates between the two sections are constant. For sections with varying ratios of sediment-accumulation rates or with hiatuses of different lengths, multiple line segments with different slopes (doglegs) are needed.

Least squares methods are generally perceived as objective, efficient and reproducible, but problems occur when doglegs are needed. Choosing the number of and positions for each line introduces unavoidable subjectivity (Edwards, 1984, 1995).

Kemple and others (1995) indicated that the geometry of the LOC is not necessarily a single straight line or even a set of straight line segments, because the ratio of the sediment-accumulation rate between two sections is always

changeable. They developed a nonlinear LOC generation method termed constrained optimization. In their method, the straight line of correlation is replaced by a piecewise correlation curve. A stochastic search method known as simulated annealing (Gershenfeld, 1999) is used to estimate the optimal position of this curve. Because the simulated annealing approach optimizes the LAD and FAD positions for every taxon on both correlated sections, there are a large number of unknown controlling parameters and associated redundant computation. As a result, the method is computationally intensive and correspondingly slow.

This paper describes an alternative correlation method using genetic algorithms with much simpler objectives. The major objective is to optimize the trend of LOC instead of the location of each datum. Since the trend of LOC is characterized by a much smaller number of parameters, our method is computationally efficient. Tests using real and hypothetical data sets show significant advantages compared with previous approaches.

Genetic Algorithms

As quantitative problems become more complex, such as by the addition of multiple independent parameters, finding the optimal solution by analytical methods becomes unfeasible (Gershenfeld, 1999; Sadler and Cooper, 2003). One well-know example of this is known as the traveling salesman problem (Papadimitriou, 1993).

The traveling salesman problem (TSP) asks that if there are N cities, with different distances apart from each other, which itinerary allows all of them to be visited with the least possible total distance traveled? Simple combinatorics show that the number of possible alternative itineraries is $0.5 \times N!$. For relatively few cities, it is simple to examine all possible solutions; e.g., for four cities there are only twelve possibilities. As the number of cities increases, however, the number of possible solutions increases dramatically. For 25 cities, there are 7.8×10^{24} possible itineraries that would need to be examined. This is well beyond the capabilities of even the fastest computers to solve in a reasonable time. What is needed, therefore, is a method for efficiently searching the multidimensional space of possible solutions and locating if not the optimal solution, then one that is acceptably close (Gershenfeld, 1999). One such approach is genetic algorithms, a branch of artificial intelligence that solves complex optimization problems by explicitly imitating neo-Darwinian evolution.

The basic approach of genetic algorithms can be described in the language of microevolution and the traveling salesman problem:

(1) A particular combination of values of the parameters, for example a specific itinerary, is digitally coded. This can be thought of as a "genotype," where the binary code of the computer replaces the base pairs of DNA in the

genetic code. The values of the parameters also can be thought of as coordinates in a multidimensional solution space.

(2) Each combination of parameter values has a corresponding solution value, such as the total distance traveled. This can be considered a measure of the "fitness" of the particular solution, in that sets of values that produce a shorter path are closer to the desired optimum and are, thus, more "fit." The particular metric used to measure fitness, the fitness function, is problem specific.

(3) An initial "population" of solutions is randomly generated and the fitness of each is determined. This population is assumed to be randomly distributed throughout the solution space.

(4) The least fit digital genotypes are removed; i.e., the longest itineraries are eliminated by "selection."

(5) The remaining, fitter solutions now "reproduce." The ones with the highest fitness can produce identical offspring. The surviving solutions also produce offspring in a manner akin to sexual reproduction, in which random crossovers and recombinations occur. This is accomplished by successful solution sets exchanging portions. There is a also provision for "mutations," in which one part of the solution is randomly changed. The purpose of recombination, crossover, and mutation is to increase the variability in the solutions and thus prevent the search from becoming stuck on a local, rather than a global optimum.

(6) The processes of selection and reproduction are repeated iteratively, until the fitness of the best solution either fails to increase or reaches some predetermined satisfactory level.

Genetic algorithms provide a powerful mechanism for searching through large solution spaces for optimal solutions. GA's resemble simulated annealing in that both are stochastic search methods. They can be used to solve a wide variety of optimization problems and do not require a thorough understanding of the problem or of optimization theory. Very few GA applications have been seen in the field of geology and have usually been used to solve geophysical inversion problems (Chunduru, 1995; Kennett and Sambridge, 1992). Bornholdt and others (1999) applied genetic algorithms to inverse stratigraphic modeling, in order to estimate the optimum fit between model parameters and an example data set.

USING GENETIC ALGORITHMS TO FIT THE LOC

Just like the TSP, fitting the line of correlation is relatively simple when only one or two line segments are needed. As the number of segments increases, however, it is much more difficult to construct the LOC, since so many

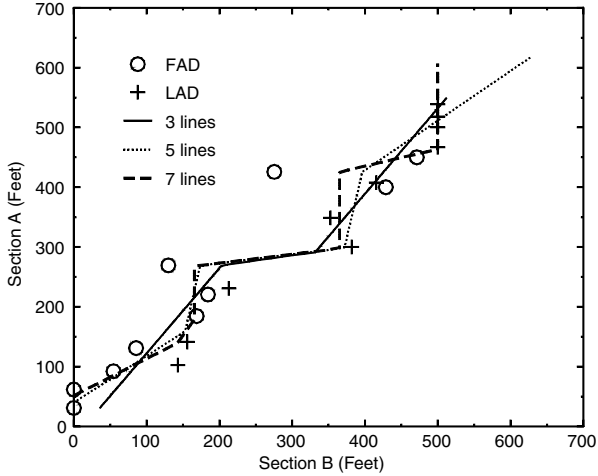

Figure 2. Three, five, and seven line segment estimates of the line of correlation through the example data set. Each LOC represents the fittest solution for that number of line segments produced by the genetic algorithm. Compare with Figure 4B in MacLeod and Sadler (1995).

alternate choices of line position and number exist (Edwards, 1995). Fitting the LOC has the additional problem that multiple approaches are available for fitting individual line segments (MacLeod and Sadler, 1995). Our approach, herein termed genetic correlation (GC), integrates genetic algorithms with classic least squares linear regression. Least squares regression has the advantage of well-established tests for statistical significance and goodness-of-fit (MacLeod and Sadler, 1995).

The fundamental concepts underlying the approach are (Fig. 2):

(1) If n is the number of datums, the LOC can be represented as a series of straight line segments ranging in number from 1 to $n - 1$.
(2) Each line has pair of starting and ending x and y coordinates.
(3) The starting x-coordinate x_1 for the first line is constrained to start at the origin; i.e., at the base of the standard section.
(4) The end x-coordinate x_{n+1} for the last line is at the top of the standard section.
(5) Ending x, y coordinates of one line are the starting coordinates for the next line.
(6) Since the first and last end points are fixed, the independent parameters are the $n - 1$ intermediate endpoint x-values. For example, if there are three

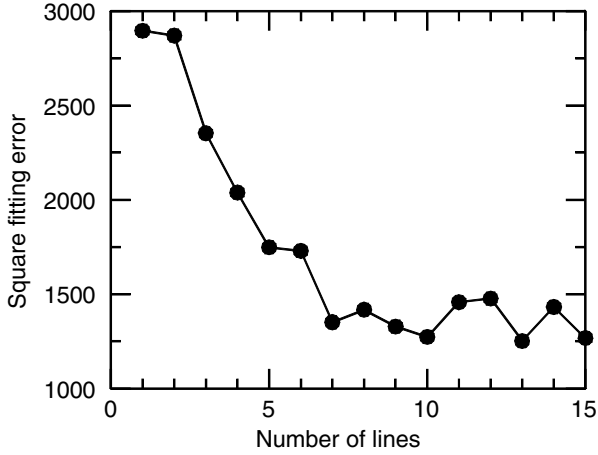

Figure 3. Fitting error as function of the number of line segments used. Although fitting error reaches a minimum when 13 lines are used, fitting quality shows little improvement when more than 7 lines are used.

lines, then there are only two adjustable values, the endpoint of the first line x_2 and the endpoint of the second line x_3 (Fig. 2).

(7) The slope is constrained to be positive or zero.

The goal, therefore, is to find the set of intermediate endpoints that produce the best fit of the total LOC to the data points.

The procedure is illustrated in Figures 2 and 3 and Table 1. The data used were taken from Miller (1977), consisting of first and last appearance datums for

Table 1. Initial (First Generation) Solution Sets for a Three Line Segment Line of Correlation for the Data Shown in Figure 1

Solution	x_2	x_3	Fitting error	Fitness rank
1	100	300	3500	3
2	200	330	2400	1
3	150	350	4000	4
4	80	400	4500	5
5	180	350	3000	2

Note. The x-coordinates of the endpoints of the first and second line segments are the independent variables x_1 and x_2. Fitting errors are the summed squared residuals of the datums from the line segments. Solutions are ranked by their relative fit; worst two fits are eliminated and do not reproduce.

Table 2. Descendent (Second Generation) Solution Sets

Solution	x_2	x_3
6	202	320
2	200	330
7	100	350
8	180	300
9	100	350

Note. Best fit from previous generation left intact. Solution 6 results from mutation of solution 2. Solution 7 is from mating of solutions 1 and 3, and solutions 8 and 9 from mating and crossover of solutions 4 and 5.

twelve unidentified taxa. The same data set was used by Kemple and others (1995) and MacLeod and Sadler (1995) and is used here to facilitate comparisons.

(1) An initial number of lines is chosen; e.g, three.
(2) A large set (up to hundreds) of possible combinations of the intermediate end points are randomly produced. For example, five possible (x_2, x_3) pairs for the data in Figure 2 are given in Table 1.
(3) Use least squares regression between x_1 and x_2 to find the slope of the first line segment and the value of y_2. The next line segment is now determined, again using regression, between the point (x_2, y_2) and x_3. This is continued until all the segments are determined.
(4) The relative fitness of a particular solution is measured by the goodness-of-fit of the series of line segments to the data. This goodness-of-fit is the sum of the squared residuals of the individual regression, so that better solutions have lower values. Solutions are ranked by fitness and solutions with the worst fit are removed. Table 1 gives the relative fitness for the five initial sets. Solutions 3 and 4 fail to survive due to their higher fitting error.
(5) Remaining solutions "reproduce" and "mate," with mutation and recombination (Table 2). The best solution, solution 2, is kept unchanged. Solution 6 is cloned from solution 2, with a slight modification (mutation); its structure is thus similar but not identical to solution 2. Solution 7 is produced by mating of solution 1 and 5, i.e., merging x_2 from solution 1 and x_3 from solution 5. Solutions 8 and 9 are similarly generated by crossover; solutions 1 and 5 exchange their x_2 values. The relative fitness of the new solutions are again evaluated and the process repeated.

(6) The process for a given number of line segments continues until the fitness metric no longer increases.
(7) Steps (2–7) are repeated with a larger number of lines; the number of lines is increased until no further improvement in fit occurs.

Figure 2 shows different multiple line segment LOC's fitted to the data of Miller (1977). Since multiple LOC geometries are almost always possible, how many lines should be preferred? Figure 3 plots the fitting error versus number of LOC lines. As expected, fitting error generally decreases when more lines are used. Although the minimum fitting error is reached when the number of line segments is thirteen, there is no noticeable improvement once the number of segments reaches seven; in fact, there is a slight decrease from seven to eight. This suggests that the increase in the number of lines used should stop when there is no net improvement in fit.

GC is an extremely efficient way to estimate the LOC. The number of parameters to be solved is an order of magnitude less than those used in simulated annealing (Kemple and others, 1995). As a result, solutions are rapidly obtained. Typical running times were less than 1 min with a 33 MHZ processor on a 1992 vintage UNIX workstation. Even shorter running time should be expected with faster operating systems.

Weighting the Fit

In the case above, all biostratigraphic datums are weighted equally. It would be logical, however, for the estimated LOC to be constrained to be more strongly controlled by points having higher reliability (MacLeod and Sadler, 1995). For example, as discussed by MacLeod and Sadler (1995), key beds with clear chronostratigraphic significance (e.g., bases of bentonite layers) should have greater value in correlation and thus more influence on the position of the LOC. Similarly, Kemple and others (1995) argued that FADs and LADs should be weighted differentially, since they are not equally susceptible to degradation by processes such as contamination or reworking. As discussed in detail by Holland and Patzkowsky (2002), range offsets, the age differences between true origination and extinction and the corresponding first and last occurrences in a section, can be on the order of millions of years. The magnitude of these offsets is strongly controlled by sequence architecture, such as the existence of major unconformities, and by ecological breadth of the species, with more stenotopic species being subject to greater offsets.

Our method can accommodate these concepts by unequal weighting of input datums in determining the regression and thus the goodness-of-fit. In weighted regression, the input values are multiplied by an *a priori* measure of their relative reliability (MacLeod and Sadler, 1995). The greater the weighting value, the greater the influence the datum has on the regression (see Davis, 2000, p. 224–228

for a fuller discussion of weighted regression). Note that although weighting may be desirable, it also introduces some degree of subjectivity; different workers may assign different weights to the same datum. Different weighting schemes lead to different LOC geometries.

It is usually assumed that confidence limits on first and last occurrences are an inverse function of the number of occurrences of a species within its range (Shaw, 1964; Signor and Lipps, 1982; Strauss and Sadler, 1989; Marshall, 1994; Holland, 1995). The relative number of occurrences in a section can thus be used as the basis of a simple linear weighting scheme. For example, assume taxa A and B both occur over the same length of section, but A occurs 4 times more commonly than B. The FAD and LAD for A thus would weight four times as heavily as those of B in determining the position of the LOC.

Numerous additional factors could affect reliability, however, such as sampling and taphonomic factors. It is important, therefore, that new objective ways to measure data quality of biostratigraphic datums be developed. These methods need to incorporate independent evidence, such as the sequence stratigraphy of the section and facies specificity of the species (Holland and Patzkowsky, 2002).

Implementation

The PGAPack genetic algorithm library, developed by David Levine (Levine, 1996) of Argonne National Laboratory was used to perform the analyses. This package can be downloaded for free from ftp.mcs.anl.gov. The package is written in C and runs on UNIX compatible workstations. Our analyses were performed on a Silicon Graphics Indigo workstation.

Applying PGAPack is straightforward. In general, the minimum requirement for the user is to specify what parameters need to be optimized (in our case, start and end points of each line segments) and what selection rule should be applied. Using PGAPack requires choosing a data structure to represent the optimization problem and specifying an objective function. It is then necessary to write a driver to integrate problem specific information with the PGAPack library functions. An example of the driver code, as well as test data sets, is available from the authors.

The performance of GC depends on a number of GA parameter settings, including population size and probabilities of selection, mutation and recombination. Larger populations will give more opportunities for finding a better solution, but also increase computational expense. Population sizes in the range of 50–150 provide a good balance of these goals. For our analyses, we used a population size of 100.

There are no clear guidelines for the values for mutation, selection, and recombination. Too high a level of mutation, for example, essentially randomizes the search. Similarly, too rigorous a selection criterion will eliminate near-optimal

models. We experimented with a variety of settings; in general we used the PGA-Pack default values. For mutation rate, this is the reciprocal of the string length (i.e., 1/(number of lines + 1)). The probability of crossover was set at 0.85. The selection criteria eliminated the 50 worst solutions at each round. Our experiments have shown that the final LOC geometries are replicable from one run to another, for a given number of lines.

Our program is configuration file driven. User can specify the sequence of running GA correlation and number of rounds by providing a configuration table. Each row of the table indicates (1) current composite section name, (2) local section name to be used, and (3) the new composite section name after merge. Our GA program will read one row at one time, execute GA correlation procedure, update composite section, save results. It then executes next row until it reaches the end of the table. In such way, the user can control project order and number of repetitions. Detailed results for LOC plot data, updated composite section data can be logged to file for analysis.

Example: Cambrian Trilobite Biostratigraphy

One the best known biostratigraphic data sets is Palmer's (1954) summary of the ranges of 62 trilobite taxa in 7 measured sections in the Cambrian Riley Formation of Texas. Shaw (1964) used the same data to introduce graphic correlation and it is frequently used as a test case for other approaches (e.g., Sadler and Cooper, 2003).

Genetic correlation is employed here to correlate these seven sections and compound them into a composite reference section (CSR). Following the same correlation sequence suggested by Shaw, the Morgan Creek section is selected as a reference section since it is the most complete. The first composite section is then formed by projecting the White Creek section onto the Morgan section. Following Shaw, who ordered the sections by their species diversity, the sections at James River, Little Llano River, Lion Mountain, Pontotoc and Streeter are then, in turn, projected onto the developing CS. As discussed above, the FAD's and LAD's were weighted by the relative abundance of each species. Once the data are entered, the correlating procedure is purely automatic and reproducible.

Figure 4 shows the LOC determined for the Streeter section plotted against the CSR. It demonstrates the non-linearity of the LOC, which reflects detailed changes of sedimentation ratio, in agreement with Kemple and others' (1995) concept of a "snake of correlation."

Validating the Method Using Stratigraphic Simulations

Figure 5 compares the Cambrian trilobite range charts obtained using GC and Shaw's Graphic Correlation method (Shaw, 1964). Kemple and others (1995) also

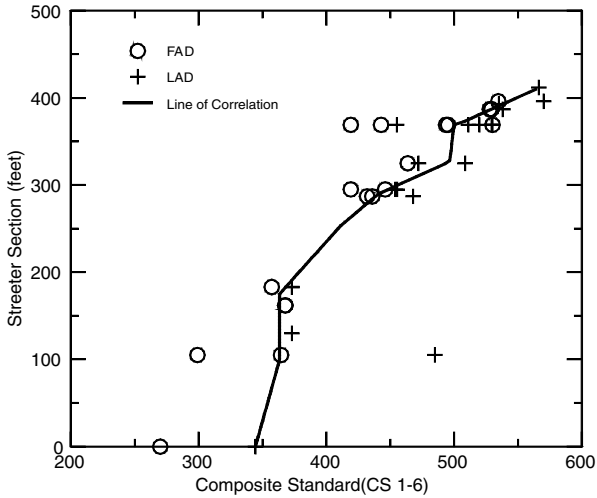

Figure 4. For the data of Palmer (1954), graphic correlation between the Streeter section and the composite standard (CS 1-6) produced by genetic correlation.

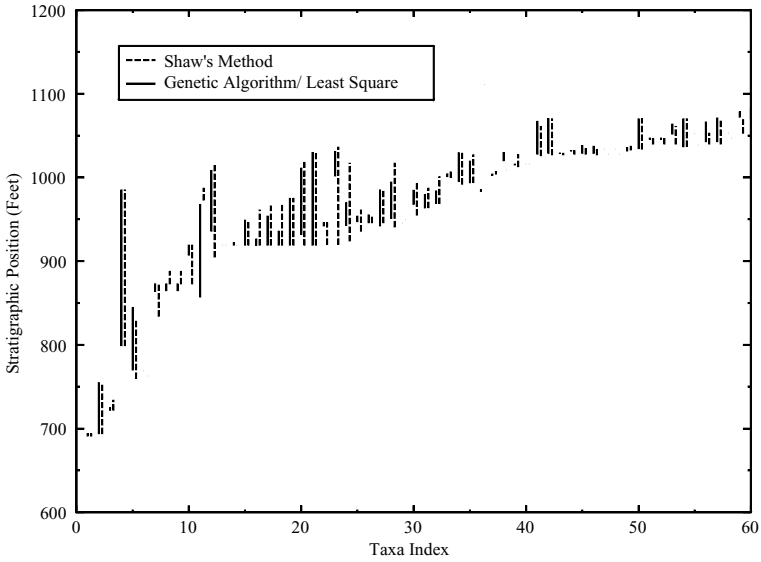

Figure 5. Trilobite range charts produced using genetic correlation compared with that produced by a conventional least squares fit.

compared their solution, obtained using simulated annealing, with Shaw's original result. All of these methods are in general agreement. Unfortunately, because the "true" range chart for these taxa is unknowable, the results from different methods can only be compared with each other, rather than to a known standard.

Simulation modeling offers an alternative approach to validating GC methodology. A basin simulation model can be used to generate synthetic biostratigraphic sections, each of which preserves an incomplete sample of the total stratigraphic ranges of an idealized group of taxa. GC, and other, can then be evaluated by the fidelity with which they reconstruct the original ranges.

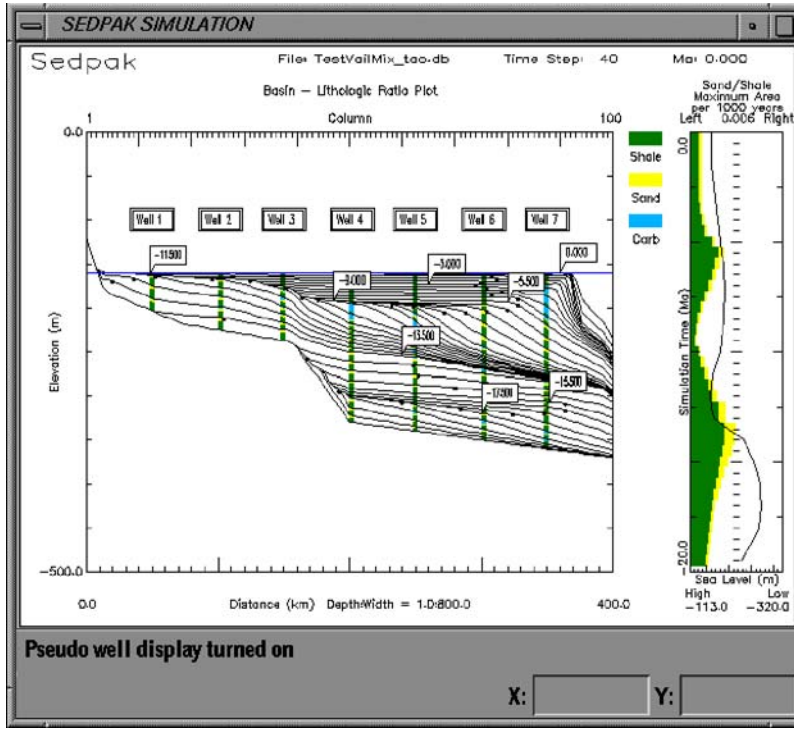

Figure 6. A hypothetical basin simulated by SEDPAK with first order sea level change (based on VailMix simulation of Cannon and others, 1994). This basin is contemporaneous with planktonic foraminifera range chart shown in Figure 7. Time span is from 20 mya to present. Initial basin surface consists of shelf, slope and continental rise Sea level (*solid line* on right side graph) shows first order changes. Sediment supply also varies with time (*solid area* on right side graph). Seven pseudo wells are drilled, ranging in depth from from shelf to slope (*vertical lines*). Thin *black lines* on cross-section are time lines; *arrowed boxes* identify particular time surfaces.

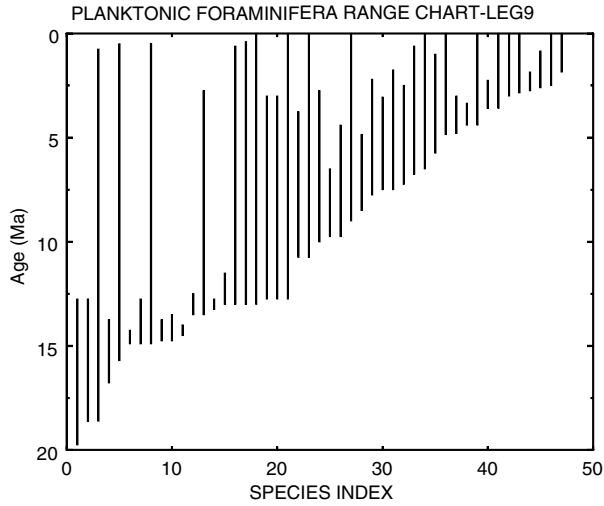

Figure 7. Ranges of 48 planktonic foraminifera taxa from early Miocene to Pleistocene (Goll, 1972). The data set is used as benchmark to validate the genetic correlation method.

SEDPAK (Strobel and others, 1989; Kendall and others, 1991) is a two-dimensional forward simulation method. It is one of many existing models used to visualize and analyze sequence stratigraphic geometries (Harbaugh and others, 1999). It assumes that stratigraphic geometry is controlled by the interplay of sediment input, sea-level change, and tectonics. The output geometry of a specific run is determined by a set of user-defined input parameters and variables. For details of the model, see Cannon and others (1994).

SEDPAK was used to create a hypothetical 400 km-wide basin which evolved over a time span of 20 m.y. (Fig. 6). The hypothetical basin has two cycles of sea level change and sediment input. The simulation used is here is VailMix simulation, which is an example supplied with the SEDPAK package (Cannon and others, 1994).

An existing planktonic foraminifera range chart, taken from the Initial Report of DSDP Leg 9 (Goll, 1972), was used as "truth" (Fig. 7). It consists of 48 taxon ranging in age from Early Miocene to Recent; i.e., the same time period as the simulated basin. For simplicity, these taxa are assumed to live in the basin and to be preserved wherever sediment of appropriate age are deposited; i.e. ranges are incomplete only because of stratigraphic hiatuses. We recognize that this is a best-case scenario and ignores ecological and other controls on occurrences.

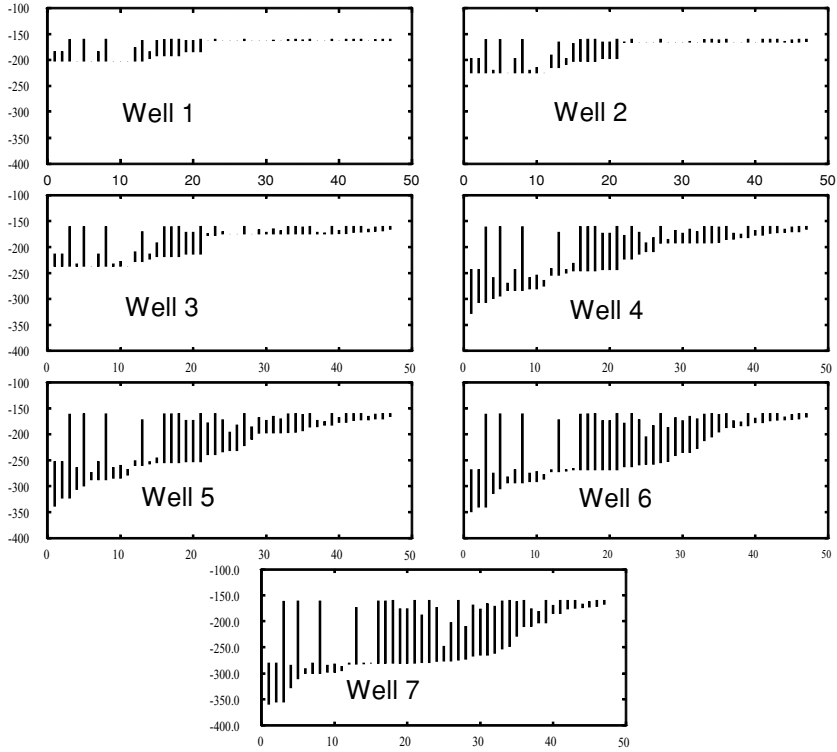

Figure 8. Foraminifera range charts for each of the seven pseudo wells along basin profile. The *x*-axis is taxon index, and *y*-axis is the depth in meters in the well.

Seven "pseudo wells" were placed along the basin cross-section at intervals of 50 km. These represent local stratigraphic sequences at these locations. The biostratigraphic sequences for each well are shown in Figure 8; they represent how the original sequence may be preserved in a series of onshore-offshore sections.

SEDPAK can produce chronostratigraphic (Wheeler) diagrams which shows periods of deposition and hiatuses as a function of geographic position and time. This allows the assessment of relative completeness of each pseudowell. Based on this, Pseudowell #6, which was the most chonostratigraphically complete, was chosen as the standard reference section (SRS).

CS was then used to reconstruct the original foraminiferal ranges. The composite range chart obtained from the synthetic sections is shown in Figure 9; it visually matches with the "truth" very well (Fig. 8). The Spearman rank correlation coefficient between the sequences is 1.00; i.e., the temporal order of the datums is perfectly recaptured.

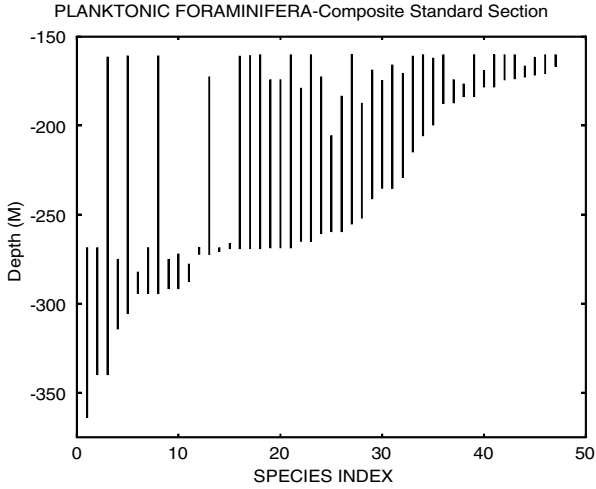

Figure 9. The composite range chart for the hypothetical basin. It is derived by projecting the other six wells onto the reference section. Compare with Figure 6.

SUMMARY AND CONCLUSIONS

We have developed a new method for fitting the line of correlation that dramatically reduces the effort involved estimating the position of the LOC and produces stable biostratigraphic correlations and composite range charts objectively and efficiently. This new method, genetic correlation (GC), is based on the use of genetic algorithms, an artificial intelligence technique which excels in locating the optimum solution from large number of alternative choices.

By using genetic algorithms, a wide range of alternative LOC's can be rapidly evaluated and a potential optimum fit determined. By using the fitness function, it is also possible to estimate the point when no further refinement of the fit is necessary. The use of genetic algorithms to estimate the LOC is estimated to be an order of magnitude more efficient than simulated annealing, an alternative artificial intelligence method (Kemple, 1995).

A major advantage of the approach described here is its flexibility. It is simply necessary to chose an appropriate method for fitting the line and a corresponding numeric fitness criterion. For example, although we used least squares regression because of the ease of calculating goodness-of-fit, the approach can certainly be adopted to other methods of line fitting, such as reduced major axis, which may be preferable due to the existence of error in both X and Y. Alternatively, the method can be used to quantify the qualitative procedure of "splitting tops and bases" (Miller, 1977; MacLeod and Sadler, 1995). This method chooses an LOC which

minimizes the number of FAD's and LAD's that are incorrectly placed relative to the standard reference section; i.e., the number of tops on the left and bottoms on the right of the LOC (Miller, 1977).

We are also investigating potential improvements for graphic correlation with the power of artificial intelligence. One of them will be an independent correlation procedure, a way to estimate the final composite standard that is independent of the order in which the sections are being incorporated. We are also planning to experiment with other line fitting algorithms, including reduced major axis. Another research area will be simultaneous correlation, an approach to fitting the line of correlation in a projection space of multiple dimensions. Finally, more experiments need to be designed to test robustness of our approach. For instance, it may be valuable to study how the LOC changes in correspondence with local terrace changes.

With exponentially increasing of computer power, and more importantly, with the increasing interest in computer modeling, the time is ripe for the broad introduction of artificial intelligence techniques to the broader geological community. Nevertheless, it must be remembered artificial intelligence techniques are tools. They do not replace, but simply supplement, the knowledge and experience of the geologists using them.

ACKNOWLEDGMENTS

This paper is based in part on Zhang's PhD thesis at the University of Illinois at Chicago. He sincerely thanks his committee members Carol Stein, Fabien Kenig, Martin Perlmutter and Torbjörn Törnqvist for their advice and assistance. Steve Holland and Peter Sadler extensively commented on an earlier version of this paper; we thank them for their efforts and assistance. Partially supported by NSF Grant EAR-9506639 to Plotnick.

REFERENCES

Agterberg, F. P., 1990, Automated Stratigraphic Correlation: Elsevier, Amsterdam, 424 p.

Agterberg, F. P., and Nel, L. D., 1982a, Algorithms for the ranking of stratigraphic events: Comput. Geosci., v. 8, no. 1, p. 69–90.

Agterberg, F. P., and Nel, L. D., 1982b, Algorithms for the scaling of stratigraphic events: Comput. Geosci., v. 8, no. 2, p. 163–189.

Bornholdt, S., Nordlund, U., and Westphal., H. 1999, Inverse stratigraphic modeling using genetic algorithms: *in* Harbaugh, J. W., Watney, W. L., Rankey, E. C., Slingerland, R., Goldstein, R., and Franseen, E., eds., Numerical Experiments in Stratigraphy: Recent Advance in Stratigraphic and Sedimentologic Computer Simulations: SEPM Special Publication No. 62, p. 85–90.

Cannon, R. L., Kendall, C. G., Levine, P., Moore, P., Ryan, S., and Wong, M. L., 1994, SEDPAK 4.0 Manual: University of South Carolina Stratigraphic Modeling Group, Columbia, South Carolina, 180 p. (also available at http://doc.igm.bo.cnr.it/sedpak/Begin.html).

Carney, J. L., and Pierce, R. W., 1999, Graphic correlation and composite standard databases as tools for the exploration biostratigrapher: *in* Mann, K. O., and Lane, H. R., eds., Graphic Correlation: SEPM Special Publication No. 53, p. 23–44.

Chunduru, R. K., 1995, Non-linear inversion of resistivity profiling data for some regular geometrical bodies: Geophysical Prospecting, v. 43, no. 8, p. 979–1003.

Davis, J. C., 2002, Statistics and Data Analysis in Geology, 3rd ed.: John Wiley & Sons, New York, 638 p.

Edwards, L. E., 1978, Range charts and no-space graphs: Comput. Geosci., v. 4, no. 3, p. 247–255.

Edwards, L. E., 1984, Insight on why graphic correlation works: J. Geol., v. 92, no. 5, p. 583–587.

Edwards, L. E., 1995, Graphic correlation: some guidelines on theory and practice and how they related to reality, *in* Mann, K. O., and Lane, H. R., eds., Graphic Correlation: SEPM Special Publication No. 53, p. 45–50.

Gershenfeld, N., 1999, The Nature of Mathematical Modeling: Cambridge University Press, Cambridge, 344 p.

Goll, R. M., 1972, Leg 9 synthesis, Radiolaria: Initial Reports of the Deep Sea Drilling Project; v. 9, p. 947–1058.

Grimm, E. C., 1987, CONISS: A Fortran 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares: Comput. Geosci., v. 13, no. 1, p. 13–35.

Harbaugh, J. W., Watney, L. W., Rankey, E. C., Slingerland, R., Goldstein, R. H., and Franseen, B. H., eds., 1999, Numerical Experiments in Stratigraphy: Recent Advances in Stratigraphic and Sedimentologic Computer Simulations: SEPM Special Publication No. 63, 362 p.

Holland, J. H., 1992, Genetic algorithms: Scientific American, v. 267, no. 1, p. 66–72.

Holland, S. M., 1995, The stratigraphic distribution of fossils: Paleobiology, v. 21, no. 1, p. 92–109.

Holland, S. M., and Patzkowsky, M. E., 2002, Stratigraphic variation in the timing of first and last occurrences: Palaios, v. 17, no. 2, p. 134–146.

Kemple, W. G., Sadler, P. M., Strauss, D. J., 1995, Extending graphic correlation to many dimensions: stratigraphic correlation as constrained optimization, *in* Mann, K. O., and Lane, H. R., eds., Graphic Correlation: SEPM Special Publication No. 53, p. 65–82.

Kendall, C. G., Moore, P. D., Strobel, J., Cannon, R., Perlmutter, M., Bezdek, J., and Biswas, G., 1991, Simulation of the sedimentary fill of basins, *in* Franseen, E. K., Watney, W. L., Kendall, C. G., and Ross, W., eds., Sedimentary Modeling: Computer Simulation and Methods for Improved Parameter Definition: Kansas Geological Survey, Bulletin 233, p. 9–30.

Kennett, B. L. N., and Sambridge, M. S., 1992, Earthquake location- genetic algorithm for teleseisms: Phys. Earth Planetary Interiors, v. 75, no. 1–3, p. 103–110.

Levine, D., 1996, Users Guide to the PGAPack Parallel Genetic Algorithm Library: Argonne National Laboratory, Publication ANL 95/18, 73 p.

MacLeod, N., and Sadler, P., 1995, Estimating the line of correlation, *in* Mann, K. O., and Lane, H. R., eds., Graphic Correlation: SEPM Special Publication No. 53, p. 51–64.

Mann, K. O., and Lane, H. R., eds., 1995. Graphic Correlation: Society of Economic Paleontologists and Mineralogists, Special Publication No. 53, 263 p.

Marshall, C. R., 1994, Confidence intervals on stratigraphic ranges: Partial relaxation of the assumption of randomly distributed fossil horizons: Paleobiology, v. 20, no. 4, p. 459–469.

Miller, F. X., 1977, The graphic correlation method in biostratigraphy, *in* Kauffman, E. G. and Hazel. J. E., eds., Concepts and Methods in Biostratigraphy: Dowden, Hutchinson, & Ross, Stoudsburg, Pennsylvania, p. 165–186.

Palmer, A. R., 1954, The faunas of the Riley Formation in central Texas: J. Paleontol., v. 28, no. 6, p. 709–786.

Papadimitriou, C. H., 1993, Computational Complexity: Addison-Wesley Publishing Company, Boston, 500 p.

Patterson, D. W., 1990, Introduction to Artificial Intelligence and Expert Systems: Prentice Hall, Englewood Cliffs, New Jersey, 440 p.

Sadler, P. M., 2004, Quantiative biostratigraphy - achieving finer resolution in global correlation.: Ann. Rev. Earth Planetary Sci., v. 32, p. 187–213.

Sadler, P. M., and Cooper, R. A., 2003. Best-fit intervals and consensus sequences: comparison of the resolving power and computer-assisted correlation, *in* Harries, P. J., ed., High-resolution Approaches in Stratigraphic Paleontology, Kluwer Academic Publishers, Dordrecht, p. 49–94.

Shaw, A. B., 1964, Time in Stratigraphy, McGraw-Hill, New York, 365 p.

Shaw, A. B., 1995, Early history of graphic correlation, *in* Mann, K. O., and Lane, H. R., eds., Graphic Correlation: SEPM Special Publication No. 53, p. 15–22.

Signor, P. W., III, and Lipps, J. H., 1982, Sampling bias, gradual extinction patterns and catastrophes in the fossil record: Geol. Soc. Am. Special Papers, v. 190, p. 291–296.

Strauss, D., and Sadler, P. M., 1989, Classical confidence-intervals and Bayesian probability estimates for ends of local taxon ranges: Math. Geol., v. 21, no. 4, p. 411–421.

Strobel, J., Cannon., R., Kendall, C. G. St. C., Biswas, G., and Bezdek, J., 1989, Interactive (SEDPAK) simulation of clastic and carbonate sediments in shelf to basin settings: Comput. Geosci., v. 15, no. 8, p. 1279–1290.

Winston, P. H., 1992, Artificial Intelligence, 3rd ed.: Addison-Wesley, Reading, Mass, 724 p.