Visual Analysis of Entity Relationships in Global Terrorism Database

Alex Godwin*, Remco Chang, Robert Kosara, William Ribarsky

University of North Carolina at Charlotte, Charlotte, NC

ABSTRACT

With the increase of terrorist activity around the world, it has become more important than ever to analyze and understand these activities over time. Although the data on terrorist activities are detailed and relevant, the complexity of the data has rendered the understanding and analysis difficult. We present a visual analytical approach to effectively identify related entities such as terrorist groups, events, locations, etc. based on a 2D layout. Our methods are based on sequence comparison from bioinformatics, modified to incorporate the element of time. By allowing the user the freedom to link entities by their activities over time, we provide a new framework for comparison of event sequences. Our scoring mechanism is robust and flexible, giving the user the flexibility to define the extent to which time is considered in aligning entities. Incorporated with high interactivity, the user can efficiently navigate through tens of thousands of records recorded in over a hundred dimensions of data by choosing combinations of categories to examine. Exploration of the terrorist activities in our system reveals relationships between entities that are not easily detectable using traditional methods.

Keywords: Visual Analytics, Terrorism, Sequence Alignment, Temporal Data

1. INTRODUCTION

We examine the Global Terrorist Database (GTD) as created by the University of Maryland's National Consortium for the Study of Terrorism and Responses to Terrorism (START) Center[1]. This dataset contains approximately 60,000 records of terrorist activities between 1970 and 1997. Approximately 120 categories are recorded for each incident, including major topics such as time, location, terrorist group responsible, weapons used, damages, etc., and detailed topics such as number of terrorists involved, type of vehicles used, number of injuries or deaths, etc.

Two specific problems are time and uncertainty in the data. Of main interest to us in this project was the comparison of the behavior of two entities (defined as a logical grouping of terrorist events, i.e., the group responsible) as it changes over time. Specifically, we were interested in assessing time periods where the behavioral patterns of one group were similar to another group. To the best of our knowledge, an analytical approach to identifying terrorist groups in the GTD based on their behaviors and patterns over time has never been done.

However, in the field of bioinformatics there has been a substantial amount of research carried out in order to determine similarity between gene sequences. While there is obviously a substantial difference between the types of data, there is a compelling similarity in the goals of sequence comparison of genes and comparison of terrorist records. In the GTD, we are attempting to find a metric for comparing the full career of a terrorist group to another as well as determining short regions of more intense similarity. In this paper, we review three pivotal sequence comparison methods presented in the annals of developed for bioinformatics research. We then present a system we have designed to apply analogous methods (albeit slightly modified) to the GTD.

In designing the visual interface, the goal was to provide a powerful view of as much of the data as possible while making it easy to see the diversity of the data and facilitate interaction. The main view is designed to provide quick and immediate comparison between two sequences and their relative alignment while giving the user the ability to easily explore the alignment in depth.

* Further author information: (Send correspondence to Alex Godwin)
A.G.: E-mail: jagodwin@uncc.edu

2. RELATED WORKS

The goal of comparing two genetic sequences in bioinformatics is most frequently to establish a phylogeny between them. If a significant comparison can be made it is an encouraging indication that the sequences have a shared ancestral relationship. An alignment between two sequences is defined as a pairwise matching of the individual components of the sequences. The following sequence alignment methods are built upon the computing principle of dynamic programming in order to drastically reduce their space and time complexity, a criterion that makes them all the more attractive for our own purposes.

What makes the algorithms below so appealing are the inherent similarities between sequences of nucleotides and the records present in the GTD. Elements of a gene sequence have a specific, immutable placement within the larger sequence, much like the occurrences for each terrorist group within the GTD. Also, a long string of nucleotides may contain the encoding for one gene or several, so multiple frames of assessment are needed as well as the ability to move between larger structures and smaller nuances of sequences. The events of the GTD, grouped by terrorist organization, represent not only the full career of each group but also several smaller event sequences worthy of comparison. Unlike gene sequences, however, the records of the GTD contain many more dimensions than a single nucleotide value. Also, the position of each nucleotide within the sequence is dictated by position within the string, not occurrence within a timeline. Nevertheless, the endeavors of bioinformatics provide a very natural starting point for us to introduce our methods, as the principles upon which our system is designed are directly related to the following algorithms.

2.1 Global Alignment: The Needleman Wunsch Algorithm

In 1970, Saul Needleman and Christian Wunsch presented a method[2] for aligning two sequences of proteins in the optimal configuration that retained all elements of both sequences in their original order. Gaps can be introduced by the algorithm to improve the alignment, and the number of gaps can be reduced by introducing a scoring penalty. The Needleman Wunsch algorithm, being the first approach to sequence alignment to employ dynamic programming, represents an important achievement in the field of bioinformatics.

Earlier attempts were blunted by the sheer number of possible alignments, but by employing dynamic programming, Needleman and Wunsch were able to bring the problem out of the realm of intractability. The immediate usefulness of a strong alignment score is that it is highly indicative of the existence of a common ancestor for the two organisms from which the sequences were obtained. The iterative matrix system of calculating the alignment, which serves as the operating principle for the Needleman Wunsch algorithm, is also the basis for several later methods of sequence alignment.

2.2 Local Alignment: The Smith Waterman Algorithm

The global alignment provided by the Needleman Wunsch algorithm, while immensely useful, can sometimes prove too rigorous for detecting shorter regions of similarity between two sequences. In order to determine the areas of concentrated similarity between two sequences, the Smith Waterman algorithm[3] provides information on short subsequences of high alignment scoring.

The Smith Waterman algorithm also employs dynamic programming, and a quick alignment can speedily provide clues to a researcher as to whether a highly scoring area in one genetic sequence is homologous to another sequence in a potential taxonomic relative. Later work has extended the usefulness of this algorithm by assessing the significance of reported sequence alignments.

2.3 Basic Local Alignment Search Tool (BLAST)

Perhaps one of the most highly cited academic papers in the 1990s, BLAST has repeatedly proven its usefulness in understanding genetic structure and inheritance since its introduction by S. Altschul et al[4]. One of the principle reasons BLAST is so successful is that it allows a user to search against a large database for potential alignments in a short amount of time. Additionally, BLAST reports statistical significance of alignments, allowing researchers to more quickly focus on making useful comparisons before drawing conclusions.

It is important to note, however, that although BLAST is guaranteed to return alignments quickly, the speed of computation comes at the price of not always returning the most optimal configuration. Nevertheless, BLAST is an incredibly useful tool for not only establishing evolutionary lines between organisms but also determining the function of unknown genes. Sequences with unknown functionality can be compared to other sequences with known functionality.

3. TIME SERIES APPLICATION

The algorithms presented in the previous section serve as an excellent foundation upon which to build, but they have to be adapted to be used for our purposes. Unlike genetic sequences, events in the real world are concretely tied to a position in time, and it is this that provides the ordering of the events for our analysis of the global terrorism database. What makes this problem particularly arduous is the difficulty in allowing time to play a meaningful role in comparison of sequences, while still uncovering alignments. The GTD spans several decades, and a researcher may not be as interested in comparing the events that differ greatly in the time in which they occurred. Our system excels at being able to allow the user to specify the exact size of the role that time will play in the structure of discovered alignments, giving researchers of the GTD the freedom to move from more global to local alignments and back again, depending on the hypothesis under scrutiny.

The comparison of two sequences of events is a logical extension of the previously discussed algorithms. Much like sequences of amino acids or nucleotides, the ordering of events during comparison must be maintained while still allowing a degree of flexibility. In this section we present the approaches we have taken to applying the methodologies of bioinformatics sequence alignment to the GTD.

3.1 Longest Common Subsequence (LCS)

The Longest Common Subsequence (LCS) algorithm is the foundation for our work in analyzing temporal sequences, and utilizes dynamic programming for reduced space and time complexity. LCS has a familiar application in comparing two files for line changes[5], and is the basis for both the Needleman Wunsch and Smith Waterman algorithms. By applying the LCS algorithm to the sequences of events for two terrorist groups, we can quickly determine the similarity of their orderings so that common sequences between the entities become visible. This is accomplished by considering each sequence of events perpetrated by a terrorist group as a string of coded values, and aligning it against a similarly coded sequence of another terrorist group.

This pattern represents an alignment between the two groups irrespective of when the events actually occurred, depending only on the sequence of events. This loose alignment allows for the detection of similar patterns of behavior even when they occur in different time periods.

3.2 Scoring Between Events

When the loose alignment provided by LCS is too general, the scoring can be modified by introducing a penalty for matches between events that are temporally distant. While the alignment derived by LCS may give an accurate and lengthy alignment between two entities, the researcher may consider them of little interest if the events occurred years or even decades apart. What's important is to allow the researcher to choose the level of constraint applied by temporal distance, view the results, and make changes as necessary. The basic match score assigned by the LCS algorithm can be modified by directly penalizing events that occur apart from one another in the timeline. Our method allows the user to define the amount of the scoring penalty and the range for which it is applied. For each range (penalties per day of difference between events, per month of difference, or per year) the amount of penalty determines the strictness of the fit.

The scoring system works by allowing the user to specify a base score for a match between events for the alignment. Once a match is determined, the current aggregate score for this alignment is increased by the match bonus. The events are compared based upon the amount of temporal distance between them. Once the amount of distance between the events is determined, the score is decreased by distance multiplied by the specified penalty.

For example, consider that an event from the first group under scrutiny occurred two years before the event from the second group. If the between-event penalty is set to 0.25 points per year of difference, then the score for the match is reduced from 1 to 0.5, and in the next iteration this event may be discarded from the alignment. In experimental trials, varying levels of alignment between entities could be quickly found by varying the between-event penalty.

By introducing a penalty for distance between events, the discovered alignment indicates a correlation between the sequences of the two groups. The reported score is provided in addition to the length of the alignment, allowing the user to determine the significance of the alignment. Early examination has been bolstered by this improvement to the sequence alignment method, as the algorithm can now return a much more focused comparison between two entities.

3.3 Gap Penalty (Local Alignment)

In order to obtain local alignments, rather than global, we introduce the ability to score penalties for gaps in the alignment. A gap is defined as a record belonging to one of the groups that is not present in the trace of the sequence alignment. The result is that our system detects smaller, tightly clustered alignments rather than the more general global alignments. By introducing a gap penalty and resetting negative scores to zero, the sequence comparison can recognize and score smaller areas of similarity. In our system the user is allowed to set the magnitude of the gap penalty; higher penalties cause the algorithm to focus on (sequentially) closer events in determining alignment.

Combining the gap penalty with the penalty between events allows the user to discover an alignment between two groups that is not only similar temporally but also concentrated locally to a subsection of each sequence of events. Analysis within the GTD has provided a validation for this method, as meaningful comparisons are readily available between groups with known associations. This is particularly true for groups that are known to have splintered from a larger organization.

3.4 Searching Against a Database

The GTD provides a vast catalog of terrorist entities to compare. Much like the BLAST system, a user specifies a sequence to compare against the other entities present in the database. In our system, the user sets the constraints for the comparison by providing the gap and between-event penalties and requesting alignments. The system defaults to local alignment when a gap penalty is not provided; otherwise the system returns the local alignment with the highest score for every other entity in the database.

Early analysis has shown promising validation when comparing terrorist entities by the targeted geographic region (e.g. Country or City) of each of their events. This alignment is particularly clear when the between event penalty is set, as the reported alignments show multiple groups that are known to be connected. Currently the system provides quick comparison for categorical dimensions, though we intend to add additional functionality in the future for comparing numerical dimensions, as well as combining the scoring of multiple dimensions simultaneously.

4. VISUAL ANALYSIS OF RESULTS

The results of an alignment are displayed in our system by allowing the user to see not only the length and placement of a determined alignment, but by also providing quick visual access to the diversity of values present in the alignment. This is an important consideration, as an alignment of two diverse sequences provides information about a more difficult to discern relationship between entities than an alignment of two relatively homogenous sequences.

4.1 Color Strip Comparison

Once a group has been selected for comparison and it has been aligned with the other entities in the GTD, the resulting alignments are displayed in a table that allows sorting by either the size of the discovered alignment or the score of that alignment (if penalties are specified by the user). The sequence alignments are displayed as a pair of color strips for the selected entity as well as each of the other entities in the database. The color strip for the entities display the position of the alignment within the sequence as well as a color value for each components of the dimension selected for comparison. Elements in the sequence that are not part of the alignment are not shaded. This visual technique allows the user to easily “trace” the path of the alignment as it winds through the sequences of the selected entity and each of the other entities in the GTD.

4.2 Trace Comparison

When a candidate sequence has been found within the results from the database search, the user can zoom in on the results of the alignment and determine the contents of the trace. A secondary view displays the elements of two groups within the trace of the alignment as well as the dates of the events for each match in the alignment. Again, a color value is displayed next to a text description of each element of the trace for easy comparison with the color strips in the primary view.

This secondary view allows the user to “drill down” into the data to determine the effectiveness of the alignment as well as formulate a hypothesis as to the nature of the alignment. It is in this view that the user may most effectively decide whether the penalties used for this alignment were effective, as well as determining whether to increase or decrease their values. This secondary view also allows the user to determine the level of uncertainty present in the matched values of the alignment, as the presented value reflects the presence of unknown values.

5. CASE STUDIES

In analysis of the GTD using our sequence alignment methods, we have validated our technique by revealing relationships between terrorist groups with a documented connection. In order to do this most effectively, a sequence of analyses is described whereby we search for the most general alignment for a group, and by showing that the visual feedback provided by the methods described in Section 3 for visualizing the results allows the user to interactively modify the parameters of the algorithm and "tighten the net," focusing on uncovering meaningful trends between entities.

5.1 Weather Underground and the Black Liberation Army

Throughout much of the 1970s, the Weather Underground was notorious for violent attacks on government buildings in the United States, particularly in New York and California. The GTD contains forty records for terrorist activities perpetrated by the Weather Underground, and we are using them as our first case study, in which we will attempt to detect shared trends based on the locations (cities) of the events. (Figure 1).

Selecting them as the primary group for sequencing and running an initial LCS analysis against the GTD (with no gap penalty or between-event penalty and a match bonus of one) reveals lengthy global alignments with the Fuerzas Armada de Liberacion Nacionale (FALN), the Jewish Defense League, and the Black Liberation Army (Figure 2). This is unsurprising given that each of the detected groups has a similar history of terrorism within the United States, also focusing largely on populated cities such as New York and Los Angeles.

Figure 1. (Left) The records of the Weather Underground are displayed to the user before analysis.
Figure 2. (Right) The user requests the LCS for each entity in the database and is presented with color strips indicating the position of the LCS within each of the other entities. Each color strip is labeled with the size of the alignment and its score.

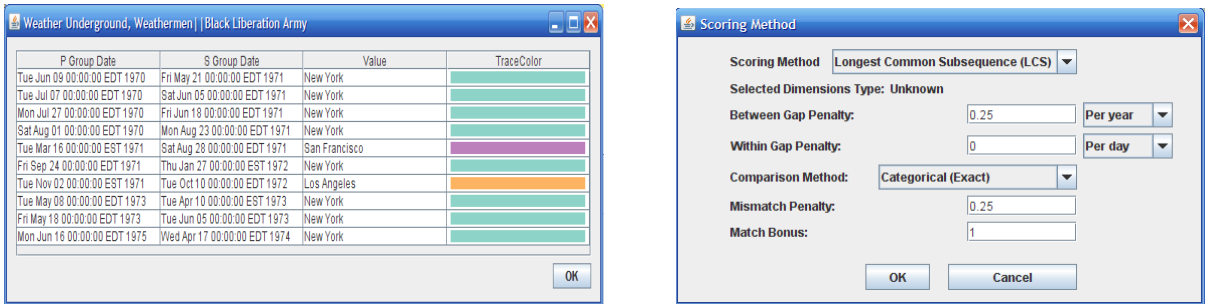


Figure 3. (Left) The elements found within a trace of the alignment of the Weather Underground and The Black Liberation Army
Figure 4. (Right) The user selects a between gap penalty of .25 points per difference in year and a gap (mismatch) penalty of .25

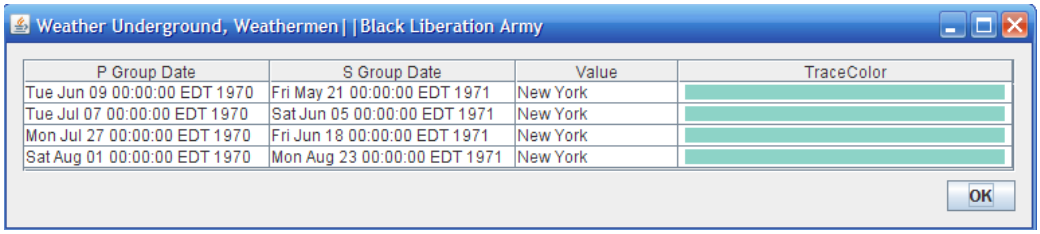


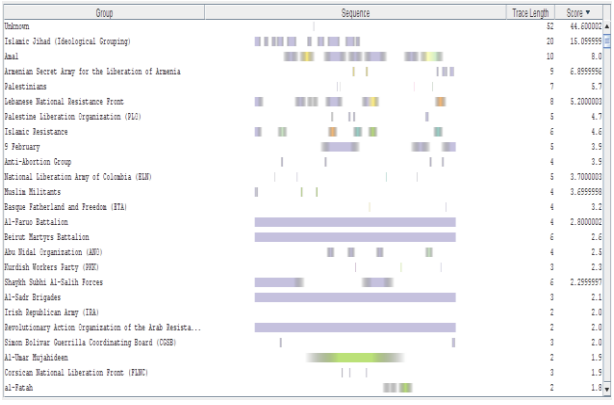
Figure 5. The elements found within a trace of the local alignment after specifying a gap penalty in addition to a time between events penalty.

In order to determine a stricter match between entities, a between-event penalty is introduced at .25 points per year of difference between the events of a match (Figure 3). Again, the results are as expected, though the Black Liberation Army is now in the highest position with an alignment trace length of 10 and a score of 9.5. The secondary view of the results shows that the majority of the elements in this alignment take place in New York, and all results are between 1970 and 1975 (Figure 4). Further evidence exists that the two groups were supportive of one another[6], and by introducing a gap penalty of .25 (forcing local alignment) we can see that each group had a sequence of four attacks in New York separated by less than a year in 1970-1971 (Figure 5).

5.2 Hizballah and the Liberation Tigers of Tamil Elaam

Hizballah, a Shi'a Islamic group based primarily in Lebanon, emerged primarily during the early 1980s during the Lebanese Civil War. When our system is tasked with finding alignments with Hizballah for the city dimension we once more find validation for our method. Initial application of a .1 per month penalty between events and no gap penalty provided several high scoring terrorist organizations from the same region (Figure 6), with those events attributed to Amal receiving a trace length of 10 and a score of 8. The events are associated with a period from 1984 to 1992 and concern fighting principally in Beirut. Amal is also based primarily in Lebanon, and most of these events are from a period marked by bitter struggle between Hizballah and Amal known as the War of the Camps[7] (Figure 7).

In order to extend the comparison, Hizballah was also examined for the "type" dimension, which describes the type of attack carried out (e.g. bombing, assassination, kidnapping, etc.) and a .1 per month penalty between events. The results showed an interesting series of alignments, with the Revolutionary Armed Forces of Colombia (trace: 171, score: 138.9) and the Liberation Tigers of Tamil Elaam (LTTE) (trace: 154, score: 124) scoring particularly high. Early periods of all three groups were marked by a mixture of bombing, assassination, kidnapping and facility attacks from the early 1908s until 1992, when all three began focusing much more on facility attacks with the occasional bombing. The decline in attack diversity is particularly striking when examining the color strips, which shift noticeably to red (Figure 8), indicating a clear shift in types of attacks for the events of both groups to facility attack.



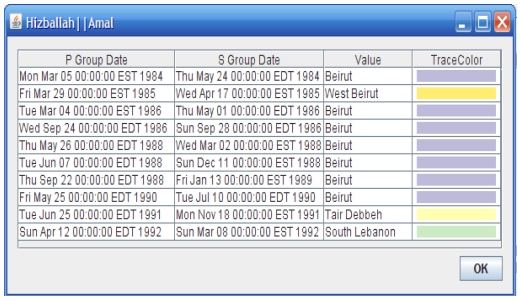


Figure 6. (Left) The user requests the LCS for each entity in the database against Hizballah for the "city" dimension using a between event penalty of .1 per month

Figure 7. (Right) The trace of the alignment of Hizballah and Amal, revealing elements from the War of the Camps and the following struggle for Beirut.

Figure 8. The user has specified a .1 per month penalty between events and requests alignments for the "type" dimension. The color strip at the top represents the placement of the traced alignment within the primary group (Hizballah). There is a noticeable shift from heterogeneity to mostly red (facility attack)

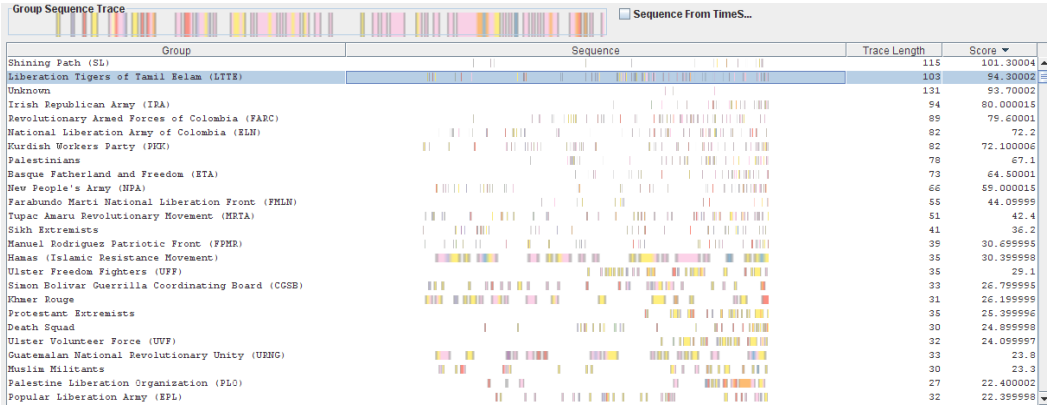


Figure 9. The user has sequenced Hizballah for the "targtype" dimension. There is a shift, though less pronounced, from heterogeneity to mostly pink (military targets)

To further explore this relationship, Hizballah was sequenced for the "targtype" dimension, which describes the nature of the target for the event (e.g. Private Citizens and Property, Government, Military, etc.) with the same .1 per month penalty between events (Figure 9). Interestingly, the LTTE (trace: 103, score: 94.3) contained once again one of the higher scoring alignments. This is particularly interesting given that the LTTE is a Sri Lankan terrorist organization with no apparent similarities to the Hizballah. What's clear is that there is a surprising similarity in methodology between the two groups worthy of further study. Close to the same time that both groups switched the type of attack primarily to facility attacks and bombing they began displaying homogeneity in type of target, choosing instead to focus on military targets with the occasional private citizen. Introducing a gap penalty of .25 per group and sequencing returns the LTTE again (trace: 25, score 11.05) as the highest scoring entity, with the local alignment focused keenly on the period from 1991 to 1992 where both groups switched target type (Figure 10).

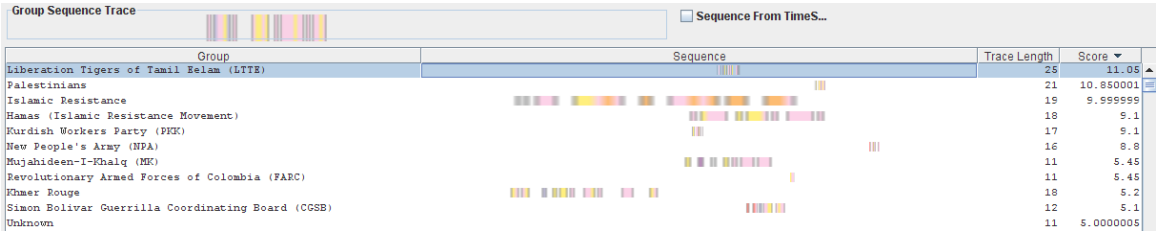


Figure 10. The user has introduced a gap penalty of .25 per mismatch. The resulting local alignment is representative of a period of transition for both the Hizballah entity and the Liberation Tigers of Tamil Eelam entity.

6. CONCLUSIONS AND FUTURE WORK

After applying these techniques to the GTD, we find that we have a reliable and fast method for comparing two terrorist groups by the ordering of their activities over time. We hope to continue our analysis of the GTD by further enhancing the ability to control the effect of time on sequence alignment, as well as furthering the level of visualization present. The initial inquiries into the data could be enhanced by providing simple descriptive views of the event sequences, as well as allowing the user to compare event sequences visually before tasking the system with an analysis query. Additionally, the nature of the LCS and our implementation is such that only the optimal alignment between two event sequences is reported, though alternate alignments exist. It would increase the effectiveness of our tool to provide the ability to the user of requesting additional information on the sub-optimal alignments and compare them.

Furthermore, the GTD contains approximately 22,000 records for which the terrorist group responsible cannot be determined. This equates to roughly one-third of the GTD. Our system currently accounts for this by defining "unknown" as an entity and allowing the user to unearth alignments between the records contained within it and other sequences. It would be particularly interesting to allow the GTD to perform several concomitant alignments for the unknown group against the other entities as a means of obtaining possible leads to the true culprit.

Additional implementation will also be carried out to allow the user to specify combinations of dimensions for which to match events, determining alignments that are the union of several factors at once. Of particular interest would be the ability to align for geographic sequences at the same time as one of the methodological dimensions, such as target type. The ability to score matches within a range, as opposed to a simple Boolean comparison, would also be of considerable worth.

The GTD is by no means the only database for which study could be enhanced by our methods. We hope to direct the use of the techniques presented in this paper to other records for which time sequence is an important consideration. For example, analysis for records of financial transactions[8] could benefit greatly from our method by allowing comparison of customer activities. This could be particularly useful in identifying fraudulent transactions, as well as classification of users by transaction sequence comparison.

ACKNOWLEDGMENTS

This work was performed with partial support from the National Visualization and Analytics Center (NVAC™), a U.S. Department of Homeland Security Program, under the auspices of the SouthEast Regional Visualization and Analytics Center. NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory.

REFERENCES

[1] LaFree, G. and Dugan, L., "Global Terrorism Database 1970-1997", [Computer file]. ICPSR04586-v1. College Park, MD: University of Maryland [producer], 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], (2007).

[2] Needleman, S. and Wunsch, C., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Biol. **48** (3), pp. 443-453, (1970).

[3] Smith, T. F. and Waterman, M. S., "Identification of Common Molecular Subsequences," J. Mol. Biol. **147**, pp. 195-197, (1981).

[4] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D., "Basic Local Alignment Search Tool," J. Mol. Biol. **215**, pp. 403-410, (1990).

[5] Hunt, J.W. and McIlroy, M. D., "An Algorithm for Differential File Comparison," *C*omputing Science Technical Report Bell Laboratories **41**, (1976).

[6] Berger, D., [Outlaws in America: The Weather Underground and the Politics of Solidarity] AK Press, pp. 177, (2006).

[7] Stork, J., "The War of the Camps, The War of the Hostages," Middle East Research and Information Project (MERIP) Reports, pp.3-7+22, (1985).

[8] Chang, R., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D., and Sudjianto, A., "WireVis: Visualization of Categorical, Time-Varying Data from Financial Transactions," Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST 2007), pp. 155, (2007).