

Chapter 2

Motif of Sequence, Motif *in* Sequence

Shin-Kap Han

What Makes a Motif a *Motif*?

Richard Coles, a radio host and chaplain of the Royal Academy of Music, recalls an edition of the weekly classical music quiz show on BBC TV, *Face the Music*, where Joseph Cooper (the chair of the show) played a single note on the piano, which Joyce Grenfell (a panel member) correctly identified as the beginning of Debussy's piano prelude *La Fille aux Cheveux de Lin* (2008). A single note!

The questions Coles poses after relating the episode are exactly the ones I would do too: How did she know? Was it a lucky guess or a photographic (or, rather, phonographic) memory on Grenfell's part? Or was there something special in the composition? If the latter were the case, what makes it a telltale signature, which at its best has the power to effectively express the whole? A better-known example is the few opening bars of the first movement of Beethoven's Fifth Symphony: Ba-ba-boom ... ba-ba-ba-boom. They seem to come out of nowhere; yet, arresting and recognizable, these few bars work with such economy that the whole of the first movement of the symphony could be described as a development of that motif.

My starting point is that this part-whole relationship—especially when the relationship is embedded in the formal structure of strings of successive states, events, actions, or notes—has a clear and close analog in the sequence analysis as practiced in social sciences in general and as implemented in optimal matching in particular. In that vein, I explore a few parallels and intersections, musical and otherwise, among these analogs to find my bearings. While advances in recent years have mostly been in methodological and technical domains, not much reflection has been seen in the theoretical domain. What I attempt here is to break the hiatus by looking outside. Throughout these excursions, the main thrust is to appropriate the concept of *motif* and its various usages in a range of extracurricular settings.

In the following section, I frame motif as a special type of subsequence and elaborate the rationale for, and the issues involved in, doing so. Three sources to

S.-K. Han (✉)

Department of Sociology, Seoul National University, Seoul, Korea

e-mail: shinkaphan@snu.ac.kr

borrow from—musical composition, molecular (or computational) biology, and social network analysis—are examined next. While disparate in substantive context, these sources provide cases that are homologous in terms of their formal structure. I search for the points of contact that will allow appropriation theoretically as well as methodologically. Finally, I conclude by gathering these threads. With a discussion of the caveats, I suggest how they can be fruitfully repurposed to advance the technical fronts and solidify the substantive bases of sequence analysis.

Local Components in Global Comparisons

The operational core of sequence analysis is to align, usually through optimal matching, two or more sequences (strings or vectors) and measure the extent to which they differ.¹ One looks for patterns shared by multiple sequences, which shed light on the structure of those sequences, and possibly what those shared patterns might do—the functions and meanings—within them. In the identification of patterns through sequence comparison, however, there are two strategies. One is *global* multiple alignment, the goal of which is to align complete sequences, and the other is *local* multiple alignment, where the aim is to locate relatively short patterns, i.e., subsequences, shared by otherwise dissimilar sequences (Lawrence et al. 1993). Thus far, a large majority of the work in social sciences follows the standard procedure of global multiple alignment (Abbott and Tsay 1990), that is, to compare and sort whole sequences. Yet, though mostly in the disciplines outside of social sciences (such as in studies of DNA), there has been continued and active interest in looking at parts of the sequences for regions of similarity or common subsequences.

This global/local distinction, however, is neither an equivalent of the level of analysis problem nor a parallel to the holism-reductionism debate one often finds in social sciences. In both strategies, the theoretical focus is on the properties in connections between the elements arrayed, such as “narrative order, sequential dependency, interlocked contingencies” (Abbott 1995) or “molecular structures and biological properties” (Lawrence et al. 1993). Both preserve the essential properties of sequence. And, more often than not, the two trek the same path in tracing the process of enchainment and unfolding. The key distinction between the two is not in *what*, but in *how*—i.e., in analytical focus. Even then, as in the case of motif discussed below, they intersect and overlap with each other.

¹ In typical practices, the result of this operation in the form of dis/similarity matrix is used for clustering, which then is presented with a variety of visualization techniques, as in Han and Moen (1999).

Table 2.1 A typology of subsequence to isolate motif

		Thematic/central in substance	
		Yes	No
Recurrent/prevalent in form	Yes	(1)	(2)
	No	(3)	(4)

Motif as a Special Type of Subsequence

Motif, in its general usage, refers to “a distinctive, significant, or salient theme or idea” or “a recurrent or prevalent characteristic” (*Oxford English Dictionary*, at “motif”). Note the pairing of the two features, one referring to the substantive aspect and the other the formal aspect of sequence-subsequence relationship. These features also key in to the two principal questions in sequence analysis: the “generating” (what produces temporal regularity?) and “pattern” questions (is there temporal regularity?), respectively (Abbott 1990; Stovel 2001).

Depending on the case at hand, however, the relative emphasis shifts between the two. In literature and literary criticism, for instance, the former is what matters most: The motif is to be elaborated, but not necessarily to be repeated.² The ‘recurrence,’ when used in these disciplines, is usually between and across works as in folklore studies (Propp 1968). In contrast, clearly apparent in music is the latter emphasis, where motif refers to a short, usually recurrent, melodic or rhythmic unit (Schenker 1980). Ravel’s *Boléro*, for instance, is a more exacting, ‘ostinato-based,’ case of such (Kamien 2010). Recurrence, not necessarily in the exactly identical form, is crucial there. In needlework and lacemaking as well as in art and architecture, it can be both, where motif may refer to a single or recurring form, shape, or color in the design or pattern (Jones 1987).

In these diverse locales, two features, either alone or in tandem, define and characterize motif as a special type of subsequence—a part that can represent the whole (*pars pro toto*). While they are not independent, they are not adjunct to each other either. When both of them are present, as in (1) of Table 2.1, it can be seen as an effective shorthand for the entire sequence and stand in for the basis for their comparisons. Even with only one of them present, as in (2) or (3), it could, though not as tidily as in (1), serve the same purpose. In these cases, and in theory, the distinction between global and local multiple alignments becomes practically moot, for the motif as a subsequence represents the complete sequence.

Presence of a motif with regard to substance, hence, suggests that there is a part, a subsequence, that contains and connotes the whole more directly—motif *of* sequence. The rest may not be as significant materially. Similarly, presence of a motif with regard to form means that there is a pattern that is being repeated—motif *in* sequence. With repetition comes redundancy, which may be dispensable. The part, of course, cannot totally contain the whole. But it always partially contains the whole

² Of interest in this context is its usage in chess, where it means an element of a move in the consideration of why the piece moves and how it supports the fulfillment of the problem stipulation. This particular usage is related to the word’s usage in French, in which motif also refers to *motive* or *purpose*.

(Kosko 1994). The more it contains in less, the better it is as a motif. If such a motif could be found, one may infer that it may be possible, so to speak, to “separate the wheat from the chaff” with proper handling. That is, instead of analyzing the entire array, one may be able to selectively focus on a part without much loss of information. Were it to be the case, thus, we gain technically in efficiency and effectiveness, for we can target more narrowly. We also gain substantively in relevance and validity, for we can concentrate on the part that matters.

On the Cutting Board

Conceptualized as a special type of subsequence, motif can be configured within the existing operational framework. As discussed below, however, doing so requires recalibrating it to focus on subsequence identification.

If s_i^k denotes the state i is in at time k , the sequence, S_i , can be represented as a vector:

$$S_i = (s_i^1, s_i^2, s_i^3, \dots, s_i^k, s_i^{k+1}, \dots, s_i^n),$$

where n , the dimension of S_i , denotes its length or the number of elements in it. S_i is aligned to the other sequences. In this alignment, the underlying premise is that these elements are not arrayed at random. Instead, some form of association (e.g., imitative, generative, etc.) between the elements is presumed, which provides the basis for a patterned regularity over time.

The alignment, though, can be done either globally or locally, comparing the sequences in whole or in part. And, at times, the distinction between the two is blurred. Take, for example, a speech that closely follows Dale Carnegie’s dictum that one should tell the audience what she is going to say, say it, and then tell them what she has said. Any of the three parts shown in Fig. 2.1-(a) can stand in for the whole, and consequently the results from the global and local multiple alignments will be identical to each other. This is a blissful case where, as in (1) of Table 2.1, a motif is found that satisfies both of the conditions.

The standard procedures for global multiple alignment are closely followed by Han and Moen in their study of the temporal patterning of retirement (1999). They obtain career pathway types by comparing respondents’ work careers and use that as a main explanatory variable. But in linking what happened before retirement to what happened during and after, they redefine the sequence boundary and analytically exploit the discrepancy between the two. In sorting career pathway types, they compare entire lengths of employment histories, thus adopting the global multiple alignment strategy. Yet when they bring in timing and type of retirement and post-retirement employment (the unshaded part in Fig. 2.1-(b)), the employment sequence (the shaded part) becomes a subsequence embedded in a lengthier sequence. In other words, the boundary of the sequence has been drawn twice, first to delineate the work career and then to extend it to retirement and beyond, thus

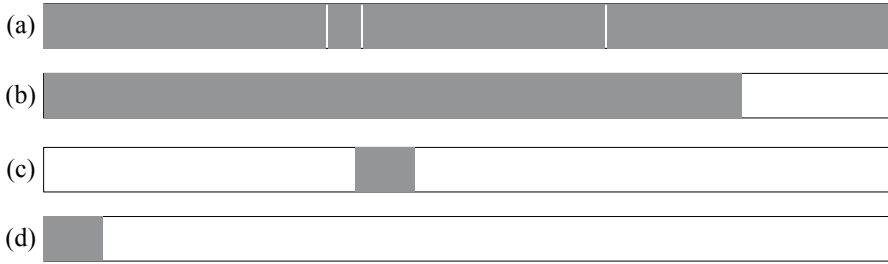


Fig. 2.1 Various sequence structures obviating the distinction between global and local alignments

turning the whole delineated at first into a part of the augmented whole later. In this setting, if we could tip the balance between the shaded (*explanans*) and unshaded (*explanandum*) parts, so much the better.

Further out along this line of reasoning lies the concept of motif. If a subsequence could be identified, which preserves and contains sequential dependency and narrative order, it could serve as an analytical catalyst. A motif as such not only obviates the distinction between global and local multiple alignments, but also provides a way to economize the process. If we let the length of a sequence be denoted by n , the sequence of dimension n has $n = \epsilon$. Thus, ϵ may take up any value within the range, $[1, n]$. In theory, that is, ϵ of a motif as a subsequence can be of a value as small as 1 and as large as n . In practice though, the smaller the value of ϵ , the better; and the larger difference, $n - \epsilon$, the bigger the gain in efficiency.³ Presented in Fig. 2.1-(c) is a hypothetical motif with its length (ϵ) less than one tenth of n .

That is, of course, if we can identify it at a search cost less than the efficiency gain accrued. The search for such a subsequence poses a difficult analytical challenge of its own. The question, in short, is how, or more specifically, *where* to cut. It involves locating the beginning and ending points of the subsequence—and hence, deciding the location and length of the subsequence—to use as a motif.

Diagrammed in Fig. 2.1-(d) is an exceptional case, in which a theory provides explicit guidance. In organizational research, “imprinting” is a concept defined as a process whereby, during a brief period of susceptibility, an entity develops characteristics that reflect prominent features of the environment, and these characteristics continue to persist despite significant environmental changes in subsequent periods. As to why organizations and industries that were founded in the same period were so similar even today, for example, Stinchcombe argues that external environmental forces powerfully shaped firms’ initial structures during the founding period, and these structures persisted in the long run, well beyond the time of founding (Stinchcombe 1965; Baron et al. 1999).⁴ This, however, is the exception that only proves

³ The gain in calculation load, which is quadratic, could be as large as $n^2 - \epsilon^2$ (Abbott and Tsay 1990).

⁴ Although Stinchcombe did not specifically use the term “imprinting,” the term soon became associated with this essay (Lounsbury and Ventresca 2002).

the rule; such clearly specified theoretical guidelines for locating subsequences would be few and far between.

In Table 2.1, the two dimensions used to classify subsequences are not equally matched in terms of feasibility. Formal recurrence is far easier to detect than substantive centrality. One does not determine the other; however, they are not entirely independent of each other either. Thus, hopefully, a better understanding of one might lead to insights on the other. With that in mind, one may look first at the formal aspect.

Resources to Mobilize

If the discussions in the preceding section are to hold water at all, we need tools to delineate a motif, a special type of subsequence. Instead of forging them anew, I look near and far to borrow. This is, after all, the strategy followed in the early period of adopting sequence analysis into sociological research (Abbott and Tsay 2000; Abbott 1995), and I am merely refreshing the process here. And I am mindful of the issues such an enterprise is fraught with, as Abbott himself noted in the following: “Specialists in these various areas may find me superficial towards their own interests even as they find me unduly concerned with those of others. These seem to me to be the inevitable costs of such a survey” (1995, p. 129). The goal here, though, is not to replicate the original materials to the letter, but to adapt them for our own, very practical use.

I set the scope of prospecting wide. Locating the structurally parallel locales, I collect appropriable analogues, harvest suitable components, and glean apposite insights and inspirations from a range of disciplines. Of course, the applicability and efficacy of these tools will depend on the setting as seen in Table 2.1 and Fig. 2.1. Yet these disparate resources, when carefully put together, might be repurposed for the problem of identifying a motif.

In the literature from sociology proper, the studies that focus on subsequence are rather limited in numbers and sources. A quick round-up will net mostly the works by Abbott: brief considerations of common subsequences in Abbott and Tsay (1990) and Abbott (1997), a discussion of “turning points” as a particular case in point in Abbott (1997), and a more explicit and elaborate treatment of it, using a Gibbs sampler, in Abbott and Barman (1983).⁵ Much of these, however, are directed toward the theoretical and substantive elaboration of “turning points,” which provides little bearing on the problem at hand.

There are, on the other hand, robust and sophisticated algorithms available, such as *TraMineR* (Gabadinho et al. 2011), to handle technical issues of subsequence identification and transition sequences. Yet their developmental tracks have been oriented mainly toward methodological and empirical purposes. Their theoretical bearings, including one on motif, have been mostly unexplored thus far. It is, in

⁵ To some extent, Hollister’s *localized* OM is based on the similar logic, i.e., on giving differential weights to different parts of the sequence (2009).

large part, due to the absence of analytical framework to articulate the two sides, which is what this chapter is after.

Looking beyond the disciplinary boundary for opportunities to borrow (and chances to steal!) yields a more interesting ensemble. I select three wide-ranging areas for such a survey below: musical composition, molecular (or computational) biology, and social network analysis.

Note for Note

I start with music. First and foremost, in its formal structure, it is a type-case of narrative that unfolds itself over time in a sequential manner (Newcomb 1987). The best example is probably the sonata form, in which a single movement is divided into three main sections: the exposition (establishing the first and second themes in contrasting keys), development (modulating in structure, tone and rhythms) and recapitulation (returning to the main themes), sometimes followed by a coda. Also, as in the earlier examples that opened this chapter, it allows a quick intuitive grasp of the notion of motif.

The focal point is musical plagiarism. To establish it in court, the plaintiff must prove that the defendant had a reasonable possibility of access to her earlier work and demonstrate that there are substantial similarities between hers and the defendant's. It is the latter problem that presents a challenge here. In theory, the outline is clear:

There must be sufficient objective similarity between the infringing work and the copyright work, or a substantial part thereof, for the former to be properly described, not necessarily as identical with, but as a reproduction or adaptation of the latter.⁶

In practice, however, it is difficult to legally define what constitutes “sufficient objective similarity.” And those difficulties keep breeding peculiarities and inconsistencies in court decisions, providing grounds for continuing legal disputes on one hand and creating needs for innovative approaches to music and law on the other.

While musical expressions involve multiple layers (e.g., rhythm, harmony, phrasing, instrumentation, and style), judging whether the two pieces share unique musical components has been done largely in terms of melodic similarities and between no more than a few measures thus far (Cronin 1998).⁷ A typical case is *Hein v. Harris* shown in Fig. 2.2.⁸ In that case, Judge Learned Hand found from his bar-by-bar analysis that thirteen of the first seventeen bars of the two melodies were “substantially the same” and concluded that Howard must have copied Hein's song.

⁶ Francis Day Hunter Ltd v Bron, Chap. 587 (1963).

⁷ We are leaving aside the issues of lyrics (e.g., Johnny Cash v Gordon Jenkins) and the recent phenomena of sampling (Vanilla Ice v Queen and David Bowie).

⁸ The composer of the defending work was Joseph E. Howard. The suit—*Hein v Harris* 175 F. 875 (C.C.S.D.N.Y. 1910)—was filed against his publisher, Harris. See Music Copyright Infringement Resource at USC Gould School of Law (mcir.usc.edu).



Silvio Hein, “Maria Cahill’s Arab Love Song”



Joseph E. Howard, “I Think I Hear a Woodpecker”

Fig. 2.2 Comparison of two melodies (Dark note heads are for unisons between the two melodies)

Another case found George Harrison liable for copyright infringement.⁹ The court’s tone is almost apologetic in determining that Harrison “subconsciously” misappropriated the “musical essence” of Ronald Mack’s “He’s So Fine” in his “My Sweet Lord.” The court relied heavily on the fact that the melodic kernels of plaintiff’s popular number were used in the same order and repetitive sequence.

In both cases, the court’s analysis seems useful and the findings essentially accurate.¹⁰ The issue of musical similarities, however, is far from settled. Some, for instance, argue that such note-for-note comparison (“by the eye”) is fundamentally incomplete given the inherent complexities of music and thus should be complemented by aural perception (“by the ear”) (Cason and Müllensiefen 2012). Still at issue too is the substantial part doctrine. Once a claimant is successful in demonstrating a sufficient degree of similarity between the disputed works, she has to establish whether the section reproduced represents a “substantial part” of the claimant’s work (Baker 1992). Note that both invoke the part-whole relationship as a principal aspect of sequence representation as discussed above. As such, these considerations prompt further questions concerning the two cases above. For the former, why the first seventeen bars, and not, say, the first eight measures? And for the latter, what are, and how does one delineate, kernels and motifs? These are the questions familiar to those who do sequence analysis in other disciplines.

Lastly, mating musical composition and computational methods engenders an interesting crossbreed—algorithmic music. It starts with the question about the very beginning: Where does music come from? David Cope, for one, believes that all music is essentially “inspired plagiarism” (2005). The great composers absorbed the music that had gone before them and their brains “recombined” melodies and phrases in distinctive, sometimes traceable, ways—a process he calls “inductive association.” He contends furthermore that such a process can be programmed. He describes this computational process in his book, *Experiments in Musical Intelligence*

⁹ *Bright Tunes Music v Harrisongs Music* 420 F. Supp. 177 (S.D.N.Y. 1976).

¹⁰ Of the existing proposals to bring in computational methods to the issue, the main part is still based on string matching algorithms that represent music as sequences of notes (Robine et al. 2007).

(1996) and presents *Emmy*, the algorithm he developed. When fed with enough of a composer's work, *Emmy* could deconstruct it, identify signature elements, and recombine them in new ways. And it actually did, producing works, including *Virtual Mozart* and *Virtual Rachmaninoff*.¹¹ Whatever this implies for the question of what music means, this logic of recombinality and the assumptions underlying it can have a profound implication for us on the ways in which sequence is seen.

The ideas and tools developed to deal with the plagiarism in music seem familiar and readily adaptable for our purpose. They might be especially useful in conceptualizing the issue of motif and the problems—both theoretical and practical—it entails (e.g., the contrast between “by the eye” and “by the ear,” “substantial part doctrine,” and “recombinality”).

A Gene is Made of DNA Sequences

As songs consist of sequences of notes, genes are made of DNA sequences. In theory, thus, what is formally true for the former is also applicable to the latter and vice versa. While, instead of melodies, protein or nucleic acid sequences are searched for shared patterns, the general outline of the approach is the same.

Upon that basis of commonality, each build their own substantive applications with distinct disciplinary orientations: In molecular biology, it is to shed light not only on molecular structure, but also on biochemical functions and evolutionary development (Lawrence et al. 1993). The early use of sequence comparison in molecular biology has largely been to detect and characterize the homology, or correspondence, between two or more related sequences, leading to evolutionary inferences. Of late, though, there has been a shift of focus in research. DNA sequence data are becoming available at a rapidly increasing rate and are now offering new ways of looking at genetic processes, including genetic diseases. After all, “[P]roteins and nucleic acids are macromolecules central to the biochemical activity of all living things, including chemical regulation and genetic determination and transmission” (Kruskal 1999, p. 3). In other words, these sequences hold the information for the construction and functioning of these organisms.

The challenge is how to read them, or perhaps more to the point, how to read more of them faster. Seizing on the fact that these sequences contain many kinds of motifs—i.e. re-occurring patterns, associated with specific biological functions, much research has been devoted to computer algorithms for discovering such motifs in sequences. These recent streams of research differ from the previous ones in three ways. One, they range much more freely across substantive domains and analytical levels, e.g., from biochemistry and neurobiology to ecology and engineering (Kashtan et al. 2004; Milo et al. 2002). They search for structural design principles across fields where complex networks constitute the structural base. In

¹¹ For those interested in how Emmy's compositional abilities fare over large numbers of compositions, he collected 5000 MIDI files of computer-created Bach-style chorales and placed at <http://artsites.ucsc.edu/faculty/cope/5000.html> for download.

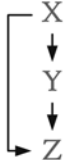

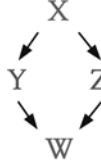
			<div>Feed-forward loop</div> <div></div>			<div>Bi-fan</div> <div></div>			<div>Bi-parallel</div> <div></div>		
Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z_{score}
Gene regulation (transcription)											
<i>E. coli</i>	424	519	40	7 \pm 3	10	203	47 \pm 12	13			
<i>S. cerevisiae</i>	685	1,052	70	11 \pm 4	14	1,812	300 \pm 40	41			
Neurons											
<i>C. elegans</i>	252	509	125	90 \pm 10	3.7	127	55 \pm 13	5.3	227	35 \pm 10	20
Food webs											
Little Rock	92	984							7,295	2,220 \pm 210	25
Chesapeake	31	67							26	5 \pm 2	8
Electronic circuits (forward logic chips)											
a15850	10,383	14,240	424	2 \pm 2	285	1,040	1 \pm 1	1,200	480	2 \pm 1	335
a38584	20,717	34,204	413	10 \pm 3	120	1,739	6 \pm 2	800	711	9 \pm 2	320

Fig. 2.3 Some network motifs found in various networks (Excerpted from Table 2.1 Network Motifs Found in Biological and Technological Networks in Milo et al. (2002, p. 826).

these works, as a result, the following three phrases are often used interchangeably: sequence motif, structural motif, and network motif. The case in point is the table in Fig. 2.3 below taken from Milo et al. (2002), which allows a fascinating comparative perspective.

Two, at the operational core of their works is counting and sampling (a small set of) subsequences and identifying those that occur at numbers that are significantly higher than those expected. Interestingly, this computational turn seems to be a return to the root of sequence analysis, i.e., string algorithms, with a “big data” twist (Sankoff and Kruskal 1999; Gusfield 1997). In this sea of DNA, the practical question is, how do we search for instances of a motif? Consequently, much of the effort in the field of late has been in accelerating the speed and expanding the scale of the search (Frith et al. 2008; Grochow and Kellis 2007).

Three, and most importantly, they go beyond morphology. They link structures with functions. In these works, network motifs are demonstrated to play key information processing roles in biological regulation networks (Shen-Orr et al. 2002). These network motifs have recently been found in diverse organisms from bacteria to humans, suggesting that they serve as basic building blocks of transcription networks (Alon 2007; Kashtan et al. 2004). Of particular interest is the “regulator gene,” involved in controlling the expression of one or more other genes.

Possibly connecting the instrumentation for sampling and counting subsequences and the analysis to specify their functions is the issue of noncoding DNA sequences.

They refer to portions of a genome sequence for which no known biochemical function has been identified, hence the label, “junk DNA” (Orgel and Crick 1980). In the human genome, for instance, more than 98% of the DNA is believed to be noncoding. If true, it could mean that it is not necessary to examine the entire length of the sequence. Increasing evidence, however, is indicating that there are discernible patterns in this noncoding DNA (Flam 1994) and it might be influencing the behavior of the coding DNA (Biémont and Vieira 2006).

The technical, especially computational, advances this field has made are certainly of interest for sequence analysts in general and particularly for those with big data implementation issues. Of more subtle, yet radical, import is the discussion of subsequences as building blocks of genes and their functions. This might as well be—if not, should be—one of the next steps for the social scientists.

Network as a Sequence, Sequence as a Network

Sequences can be conceived and represented as digraphs (or directed graphs) (Harary et al. 1965). For sequences, however, the arcs in them are restricted to be only in one direction. A type-case is chronological sequences, in which the arcs must be directed from time t to time $t + k$ ($k > 0$). Even when there is no intrinsically determined way to designate origin and destination points, one chosen direction, whether from left to right or from top to bottom, is adhered to. As such, they form vectors, i.e., one-dimensional arrays. Strings of relations can readily be adapted to the sequence framework. Networks as such are digraphs—and hence, sequences—as well. They differ, however, in that they are typically two- or three-dimensional arrays. Establishing a formal analogy between networks and sequences thus requires a dimensional shift. With that as a caveat, one may look to social network analysis (SNA) to pan for materials.¹²

In the literature, there has long been a strand that looks at the distribution of parts (or smaller structures) to make sense of the nature of the whole (the larger structure). Triad census, based on all sixteen types shown in Fig. 2.4, is an earlier example of analysis using subgraphs (Holland and Leinhardt 1970). Watts (2004) finds the work on network motifs by molecular biologists, such as (Shen-Orr et al. 2002) and Milo et al. (2002), “identical in spirit” to this literature. There seems to be far more than spiritual similarity: It is analogous in defining network motifs as topologically distinct subgraphs whose frequencies in a network can be used to characterize its overall structure. It is also technically parallel—systematically enumerating all the motifs comprising three and four nodes in a number of networks first, and then comparing the resulting counts with those from random networks.

¹² In this volume, Bison too exploits this linkage, in a direct, albeit quite different, manner, by plotting sequences as networks. Such a translation, he argues, provides an alternative way to observe and measure new structural features of complex narratives (Bison 2012; Abell 2004; Bearman and Stovel 2000).

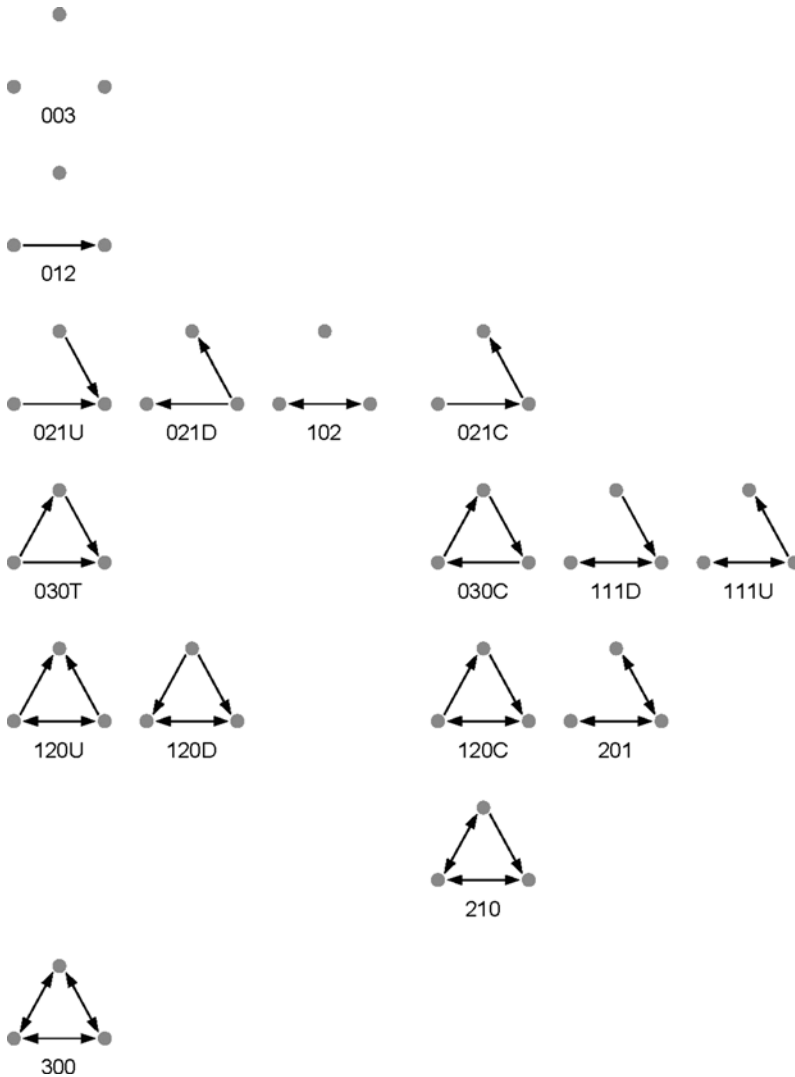


Fig. 2.4 Sixteen triad types (All sixteen triad types arranged vertically by number of choices made and divided horizontally into those with no intransitivities and those with at least one. See Fig. 2.1 in Holland Leinhardt (1970, p. 496) and Fig. 2.4 in Han (2003, p. 268)

Of more recent vintage along this line of research are the exponential-family random graph models (ERGMs) of networks. The basic stochastic model can be expressed by the following general form (Wasserman and Robins 2005):

$$Pr(Y = y) = \left(\frac{1}{k}\right) \exp\left\{\sum_A \eta_A g_A(y)\right\},$$

where Y is a network realization and y is the observed network. The summation is over all configurations A . k is a normalizing factor. η_A is the parameter and $g_A(y)$ is the network statistic corresponding to configuration A .

The purpose of ERGMs is to describe parsimoniously the local selection forces that shape the global structure of a network. That is, a social network is thought of as being built of these local patterns of ties, called “network configurations,” which correspond to the parameters in the model (Lusher et al. 2012). In this framework, for instance, one may ask, does a given network structure occur due to processes of homophily, reciprocity, transitivity, cyclicity, or a combination of these?¹³ And as such, the formulation also takes into account the inferential potential of the sequence analysis (King 2013). The efforts at substantive as well as technical developments of late are mostly directed toward inclusion of higher-order local structures, such as k -stars, k -triangles, and independent two-paths and their alternating versions.¹⁴

In this framework, we are brought back to reconsider the part-whole relationship discussed earlier in formulating the concept of motif. Yet the two sides are engaged not just formally, but in theoretical and substantive ways as well. Such a dual engagement is precisely what will allow us to consider the two dimensions in Table 2.1 simultaneously: recurrent/prevalent in form and thematic/central in substance.

Concluding Remarks

I explored a few avenues that intersect our main topic—the core program of “sequence analysis,” in which sequences are strings of successive states, events, or actions, and for which optimal matching serves as the standard operational framework. In those excursions, the main thrust was in appropriating and repurposing the concept of motif, defined as a distinctive and recurring set of structural elements, and its various usages from a range of extracurricular settings. This focus on motif, of course, is not effective everywhere. But, where it works, it could generate interesting and important leads.

Although certainly not exhaustive in any way, a few leads are found scattered in diverse settings in varying shapes. And much has been gained, hopefully, culling usable bits and pieces to advance the technical front on one hand, and solidifying the substantive embedding of sequence analysis on the other. Those diverse settings all deal with arrays of element. The elements arrayed may be informational or corporeal. And the arraying may be linear/temporal or multidimensional/spatial. Yet the analytical issues—especially, in their structural forms—are analogous, which allows borrowing from one another. The key is to see the sequences as built from,

¹³ Currently, they are implemented in *statnet* (<http://statnet.csde.washington.edu/>) and *PNet* (<http://sna.unimelb.edu.au/PNet>).

¹⁴ Appendix A. Table of Model Terms in (Morris et al. 2008) provides quick reference for what terms are appropriate to a particular model.

with, and by component blocks, chunks of elemental, basic units that form a minimal substantive footing, akin to White's *social molecules* (1992).

These examples do resemble one another, particularly in highlighting the common subsequence problem. They all see the potential for motif. And at a deeper level, they see that sequences could be generated endogenously and recursively, i.e., that certain subsequences, stages, events or specific episodes could have an influence on the further enchainment of the sequential elements, as in the phrase “defining moments” or “critical events” (Blanchard et al. 2012). Yet the fact that music and narrative, for instance, both involve a succession of events in a regular order does not necessarily mean that music has a special affinity to narrative (Maus 1991). In general, claims that two kinds of object are close in terms of formal structure are risky: it is too easy to find and describe shared structures across many different domains (Kruskal 1999). That is, while it is useful to exploit the analogies between them, it is imperative to pay attention to the significant empirical differences and theoretical distinctions between them as well to avoid superficial transfer (Biemann 2011).

Further Issues to Consider

There are a few issues to consider to implement the idea of motif in practice. One of the key issues is that of bounding the motif. One twist here is that in so doing, we have to find a halfway stop between the parts and the whole—i.e., at a subsequence level. The problem has implications on several levels.

First and foremost, the underlying presumption on the nature of sequence structure is that, empirically, the elements are *not* arrayed randomly. The patterned regularities we seek are taken as the results of that non-randomness (such as “enchainment, order, convergence” (Abbott 1995)). In turn, theoretically, we presume that there are underlying processes that produce these regularities. Within this overall framing, motif specifies that those theoretical and empirical keys are to be found in the constituent components of sequences. As discussed earlier, however, it does not explicitly specify how big, or how long, those components are (Elzinga and Wang 2012). At another level are the related issues of granularity¹⁵, gap, and nestedness, which are particularly difficult ones in temporal dimensions. Thus far, the answers to these questions have been dictated largely by the exigencies of available data without much articulation.

While these issues may seem to be of more empirical/methodological nature, they touch upon the fundamental issue of how to break things down and how to put them back up. That is, if we are to understand sequences as social narratives, we have to identify not only their temporal structures, but the interdependent processes in (of?) them as well. With motif, in particular, we have to ask: What are the funda-

¹⁵ Granularity in general is the extent to which a system is broken down into small parts, either the system itself or its description or observation.

mental building blocks? How do they combine to form larger structures? Do these structures which share the same building blocks also share the same combinations of these blocks?

Looking Backward, Looking Forward

In the invitation letter for LaCOSA conference, the organizers write: “We currently lack a broader and systematic debate that takes stock of the advances and limits of sequence analysis, that encourages a careful standardization of the approach beyond diverging orientations, and that opens and explores new methodological paths and combinations.” For that, then, let’s ask again what we do sequence analysis for. In it, the first problem is always to figure out if the patterns are there. By ‘patterns,’ we mean regularities in sequence, as in sequence types, and we look for them by comparing sequences. At various stages of this classification exercise, we continuously engage in data reduction—either by necessity or for convenience. That much is clear from the technical point of view. Imperative as those demands are, we should also keep our eyes on the prize, i.e., theorizing intertemporal dynamics. That is, after, and at times simultaneously with, the classification, we use that ‘reduced-form data’ to do explaining-modeling-theorizing about the processes of unfolding, the mechanisms of entailment, and the structures of temporal space. And that is what we should keep our eyes on. As Abbott forcefully puts it, “The proof of the classificatory pudding comes in the explanatory eating” (1990, p. 15).

In exploring this avenue, it might be helpful to take lessons from a kindred experience. In an essay aptly titled “Structural Analysis: From Method and Metaphor to Theory and Substance,” Wellman (1988, pp. 19–20) poignantly describes the predicament of social network analysis as of 1988:

These misconceptions have arisen because too many analysts and practitioners have (mis) used “structural analysis” as a mixed bag of terms and techniques. Some have hardened it into a method, whereas others have softened it into a metaphor. Many have limited the power of the approach by treating all units as if they had the same resources, all ties as if they were symmetrical, and the contents of all ties as if they were equivalent.

Yet, structural analysis does not derive its power from the partial application of this concept or that measure. It is a comprehensive paradigmatic way of taking social structure seriously by studying directly how patterns of ties allocate resources in a social system. Thus, its strength lies in its integrated application of theoretical concepts, ways of collecting and analyzing data, and a growing, cumulating body of substantive findings.

There are quite a few meta-theoretical parallels and practical similarities between social network analysis then and sequence analysis now: We are very much at that juncture, where the direction for us too is *from* method and metaphor *to* theory and substance. To pose new intellectual questions, collect new types of evidence, and provide new ways to describe and analyze social structures are what sequence analysis has to achieve. And, for that, thinking about sequence in terms of motif, and looking for the motif in social narrative, is a turn that might show a new path.

Acknowledgement This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-330-2012S1A3A2033451). I thank Sang-Jic Lee, who provided help in the manuscript preparation, and Yun-joo Sung, who prepared the scores in Fig. 2.2.

References

- Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods*, 16(4), 129–147.
- Abbott, A. (1990). A primer on sequence methods. *Organizational Science*, 1(4), 375–392.
- Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113.
- Abbott, A. (1997). On the concept of turning point. *Comparative Social Research*, 16, 85–105.
- Abbott, A., & Barman, E. (1997). Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology*, 27, 47–87.
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods and Research*, 29(1), 3–33.
- Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology*, 30, 287–310.
- Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450–461.
- Baker, M. (1992). La(w)-A note to follow so: Have we forgotten the federal rules of evidence in music plagiarism cases? *Southern California Law Review*, 65, 1583–1637.
- Baron, J. N., Burton, M. D., & Hannan, M. T. (1999). Engineering bureaucracy: The genesis of formal policies, positions, and structures in high-technology firms. *Journal of Law, Economics, and Organization*, 15(1), 1–41.
- Bearman, P. S., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics*, 27(2), 69–90.
- Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology*, 41, 195–221.
- Biémont, C., & Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. *Nature*, 443(7111), 521–524.
- Bison, I. (2012). *Sequence analysis and network analysis: An attempt to represent and study sequences by using NetDraw*. In Lausanne Conference on Sequence Analysis (LaCOSA).
- Blanchard, P., Buhlmann, F., & Gauthier, J.-A. (2012). *Sequence analysis in 2012*. In Lausanne Conference on Sequence Analysis (LaCOSA).
- Cason, R. J., & Müllensiefen, D. (2012). Singing from the same sheet: Computational melodic similarity measurement and copyright law. *International Review of Law, Computers & Technology*, 26(1), 25–36.
- Coles, R. (2008). Got it licked. The Guardian. www.guardian.co.uk/music/2008/jul/22/popandrock.classicalmusicandopera Accessed 8 May 2013.
- Cope, D. (1996). *Experiments in musical intelligence*. Middleton: A-R Editions.
- Cope, D. (2005). *Computer models of musical creativity*. Boston: The MIT Press.
- Cronin, C. (1998). Concepts of melodic similarity in music-copyright infringement suits. *Computing in Musicology*, 11, 187–209.
- Elzinga, C., & Wang, H. (2012). Versatile string kernels. In *Lausanne Conference on Sequence Analysis* (LaCOSA).
- Flam, F. (1994). Hints of a language in junk DNA. *Science*, 266(5189), 1320. PMID: 7973718.
- Frith, M. C., Saunders, N. F. W., Kobe, B., & Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, 4(5), e1000071.

- Gabardin, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with *traminer*. *Journal of Statistical Software*, 40(4), 1–37.
- Grochow, J. A., & Kellis, M. (2007). Network motif discovery using subgraph enumeration and symmetry-breaking. In T. Speed & H. Huang (Eds.), *Proceedings of the 11th annual international conference on research in computational molecular biology (RECOMB'07)* (pp. 92–106). Berlin: Springer-Verlag.
- Gusfield, D. (1997). *Algorithms on strings, trees and sequences: Computer science and computational biology*. New York: Cambridge University Press.
- Han, S.-K. (2003). Tribal regimes in academia: A comparative analysis of market structure across disciplines. *Social Networks*, 25, 251–280.
- Han, S.-K. & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology*, 105(1), 191–236.
- Harary, F., Norman, R. Z., & Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs*. New York: Wiley.
- Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, 76(3), 492–513.
- Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38(2), 235–264.
- Jones, O. (1987). *The grammar of ornament: All 100 color plates from the Folio edition of the great Victorian sourcebook of historic design*. New York: Dover Publications.
- Kamien, R. (2010). *Music: An appreciation*. New York: McGraw-Hill.
- Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11), 1746–1758.
- King, T. (2013). A framework for analysing social sequences. *Quality and Quantity*, 47(1), 167–191.
- Kosko, B. (1994). *Fuzzy thinking: The new science of fuzzy logic*. New York: Flamingo.
- Kruskal, J. B. (1999). An overview of sequence comparison. In D. Sankoff & J. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Stanford: CSLI Publications.
- Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262(5131), 208–214.
- Lounsbury, M., & Ventresca, M. J. (2002). Social structures and organizations revisited. *Research in the Sociology of Organizations*, 19, 3–26.
- Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). *Exponential random graph models for social networks: Theory, methods, and applications*. New York: Cambridge University Press.
- Maus, F. E. (1991). Music as narrative. *Indiana Theory Review*, 12, 1–34.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 824–827.
- Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, 24(4), 1548–7660.
- Newcomb, A. (1987). Schumann and late eighteenth-century narrative strategies. *Nineteenth-Century Music*, 11(2), 164–174.
- Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: The ultimate parasite. *Nature*, 284(5757), 604–607.
- Propp, V. (1968). *Morphology of the folktale*. Austin: University of Texas Press.
- Robine, M., Hanna, P., Ferraro, P., & Allali, J. (2007). *Adaption of string matching algorithms for identification of near-duplicate music documents*. In Proceedings of the International SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN'07) (pp. 37–43). Amsterdam, The Netherlands.
- Sankoff, D., & Kruskal, J. (Eds.). (1999). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Stanford: CSLI Publications.
- Schenker, H. (1980). *Harmony*. Chicago: University of Chicago Press.

- Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1), 64–68.
- Stinchcombe, A. L. (1965). Social structure and organizations. In J. G. March (Ed.), *Handbook of Organizations* (pp. 142–193). Chicago: Rand McNally & Company.
- Stovel, K. (2001). Local sequential patterns: The structure of lynching in the deep south, 1882–1930. *Social Forces*, 79(3), 843–880.
- Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p^* . In P. J. Carrington, J. Scott & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 148–161). New York: Cambridge University Press.
- Watts, D. J. (2004). The “New” science of networks. *Annual Review of Sociology*, 30(1), 243–270.
- Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz (Eds.), *Social structures: A network approach* (pp. 19–61). New York: Cambridge University Press.
- White, H. C. (1992). *Identity and control: A structural theory of social action*. Princeton: Princeton University Press.

Advances in Sequence Analysis: Theory, Method,
Applications

Blanchard, P.; Bühlmann, F.; Gauthier, J.-A. (Eds.)

2014, XIII, 304 p. 75 illus., 35 illus. in color., Hardcover

ISBN: 978-3-319-04968-7