Chapter 16
Pattern and Antipattern Discovery in Ethiopian Bagana Songs

Darrell Conklin and Stéphanie Weisser

Abstract Pattern discovery is an essential computational music analysis method for revealing intra-opus repetition and inter-opus recurrence. This chapter applies pattern discovery to a corpus of songs for the bagana, a large lyre played in Ethiopia. An important and unique aspect of this repertoire is that frequent and rare motifs have been explicitly identified and used by a master bagana teacher in Ethiopia. A new theorem for pruning of statistically under-represented patterns from the search space is used within an efficient pattern discovery algorithm. The results of the chapter show that over- and under-represented patterns can be discovered in a corpus of bagana songs, and that the method can reveal with high significance the known bagana motifs of interest.

16.1 Introduction

Pattern discovery is a central component of music analysis, concerned with the induction of patterns, those that are perceptually salient, statistically distinctive, or interesting to an analyst, within a single piece of music (*intra-opus*) or in a corpus of pieces (*inter-opus*). In contrast to a *deductive* analysis, which proceeds from known or postulated patterns and queries for their occurrences within a corpus, an *inductive* analysis performs the converse inference, proceeding from the music surface to salient patterns. The identification of salient structures within a piece or a corpus is a

Darrell Conklin
Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastián, Spain
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
e-mail: darrell.conklin@ehu.eus

Stéphanie Weisser
Université libre de Bruxelles, Brussels, Belgium
e-mail: stephanie.weisser@ulb.ac.be

Conklin, D., & Weisser, S.. (2016). Pattern and antipattern discovery in Ethiopian bagana songs. In D. Meredith, Computational Music Analysis (1st ed., pp. 425-443). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-25931-4_16

central part of musical analysis. In contrast to current methods for pattern discovery in music (Conklin and Anagnostopoulou, 2001; Lartillot, 2004; Meredith et al., 2002), which rely to varying extent on a form of salience determined by intra-opus repetition and inter-opus recurrence, a recent study (Conklin, 2013) has considered the problem of discovering *antipatterns*: patterns that are surprisingly rare or even absent from a piece or corpus. This property, which Huron (2001) refers to as a *negative presence*, has so far seen little attention in the computational music analysis literature.

Since the virtual space of patterns may be large or even infinite (in the case, for example, of antipatterns), all pattern discovery algorithms must address the dual problems of efficiently searching this space, and of ranking and presenting the results. Subsequently, a second order evaluation process arises: are patterns that are highly ranked also interesting to the analyst? Therefore it can be extremely productive to deeply study pieces where prior motivic analyses are available (Collins et al., 2014), for example as performed by Huron (2001), and Conklin (2010b) in their computational validations of Forte's (1983) motivic analysis of the Brahms String Quartet in C Minor.

In this chapter we will present a study of pattern discovery in Ethiopian bagana songs. The study of a bagana corpus provides a unique opportunity for the evaluation of pattern discovery techniques, because there are known inter-opus motifs, both rare and frequent, that are significant for didactic purposes (Weisser, 2005). The aim of this study is to explore whether inductive methods for the discovery of significantly frequent and rare patterns in music can reveal the known motifs, and possibly others that have functional significance.

Sequences are a special form of data that require specific attention with respect to alternative representations and data mining techniques. Sequential pattern mining methods (Adamo, 2001; Agrawal and Srikant, 1995; Ayres et al., 2002; Mooney and Roddick, 2013) can be used to find frequent and significant patterns in datasets of sequences, and also sequential patterns that contrast one data group against another (Deng and Zaïane, 2009; Hirao et al., 2000; Ji et al., 2007; Wang et al., 2009). In music, sequential pattern discovery methods have been used for the analysis of single pieces (Conklin, 2010b), for the analysis of a corpus of pieces (Conklin, 2010a), and also to find short patterns that can be used to classify melodies (Conklin, 2009; Lin et al., 2004; Sawada and Satoh, 2000; Shan and Kuo, 2003).

Further to standard pattern discovery methods, which search for frequent patterns satisfying minimum support thresholds (Webb, 2007), another area of interest is the discovery of rare patterns. This area includes work on rare itemset mining (Haglin and Manning, 2007) and negative association rules (Wu et al., 2004) which have seen application in bioinformatics (Artamonova et al., 2007). For sequence data, rare patterns have not seen as much attention, but are related to *unwords* (Herold et al., 2008) in genome research (i.e., absent words that are not subsequences of any other absent word), and *antipatterns* (Conklin, 2013) in music (patterns that are surprisingly rare in a corpus of music pieces). Antipatterns may represent structural constraints of a musical style and can therefore be useful for classification and generation of new pieces.

Fig. 16.1 The Ethiopian lyre *bagana*. Photos: Stéphanie Weisser

This chapter is structured as follows. Section 16.2 provides some historical context on the Ethiopian bagana, the notation used to encode notes in bagana songs, the known motifs, and a description of the encoded corpus of bagana songs. Section 16.3 outlines the pattern discovery method, statistical evaluation of patterns, and an algorithm for pattern discovery. Section 16.4 presents the computational analysis of the bagana corpus, first for antipatterns, then for positive patterns. The method is able to reveal known bagana motifs in a small corpus of bagana songs. The chapter concludes with a general discussion on the role of inductive methods in computational music analysis.

16.2 Bagana Music and Notation

In this chapter known frequent and rare motifs proposed by a master bagana teacher are studied using pattern discovery methods. Before presenting the motifs and corpus, a brief introduction of the cultural context will serve to provide the relevant background in playing technique and notation used to encode pieces.

The Ethiopian lyre *bagana* is a large instrument (ca. 1.20m high), made of wood, cattle skin and gut strings (see Fig. 16.1). It is played by the Amhara, a people settled mostly in Central and Northern Ethiopia. As a stringed instrument with strings running to a yoke or crossbar held by two arms coming out of the resonator, the bagana is categorized by organologists as a lyre. While such instruments were largely

Fig. 16.2 Bagana representations from ancient manuscripts. Left: Illumination (detail) from Gospel of St. John (fol. 3v), Debre Bankol, Shire, Tigre Province, 19th century. Photo courtesy of Michael Gervers, University of Toronto. Right: Illumination (detail) from Life of St. Täklä Haymanot & Gäbrä Mänfäs Qeddus, Story of Archangel Mikael (fol. 119v), Samuel Za-Qoyetsa Monastery, Tigre Province, 18th century. Photo courtesy of Michael Gervers, University of Toronto

played in ancient Mesopotamia, ancient Egypt and ancient Greece, millennia ago, they are nowadays only found around the Nile and in East Africa.

The bagana plays a special role in Amhara culture: it is indeed the only melodic instrument belonging to the *zema*, the Amhara spiritual musical sphere. All the other instruments of the zema are percussive, producing sounds without definite pitch. The myth of its origin might explain such a specific status: according to Amhara belief, the bagana is the biblical King David's instrument, brought to Ethiopia together with the Ark of the Covenant by Menelik I, legendary son of King Solomon and Queen of Sheba. Such close association with the divine election and with the founder of the Solomonic dynasty explains why the bagana was historically an instrument played by pious men and women of the nobility (*makwannent*).

Even though it is not played during liturgical ceremonies, the bagana is considered as a sacred instrument, whose sounds are able to create a special connection with God, the saints and the Virgin Mary. For this reason, it is the only instrument traditionally allowed to be sounded during Lent (*Fasika Tsom*), the long fasting period before Easter. The bagana is also considered as a very powerful object, able to cast away the devil and to protect a house from evil spirits. This power against evil forces is linked to the biblical text of Samuel (1 Samuel 16:14-23), recalling the episode of King David playing the instrument to appease King Saul, tormented by an evil spirit. Indeed, the bagana is often depicted in iconographic representations being played by King David in majesty on his throne (see Fig. 16.2). The bagana is therefore a

highly-respected instrument, inducing strong emotional states for players as well as for listeners (Weisser, 2012).

The bagana is a solo instrument accompanied by singing voice only and produces very low pitch sounds (from 50 to 200 Hz in general, which is close to the tessitura of a double bass). Bagana sounds are characterized by their very specific buzzing sound quality: indeed, due to a specific device named *enzira* and located at the bridge, the instrument generates extremely rich and harmonic sounds, perceived by the human ear as buzzing.

A bagana song is usually referred to as a *mezmour* (spiritual song), and it is built on repetition: a relatively short melody is repeated up to 20 times, with different lyrics each time except in the refrain (*azmatch*). Bagana songs are usually preceded by instrumental preludes, called *derdera* (pl. *derderotch*). Lyrics are an important part of the bagana repertoire, as they comprise prayers, praise to God, Jesus, the Virgin Mary and the saints, and elaborate poetry including *semenawork* (literally "wax-and-gold", a literary construct widely used by Amhara and characterized by double- or triple-meaning). Musical repetition is not strict: several modifications can be introduced by players in the course of the song, but they are usually subtle. The perceptual effect of the musical repetition, combined with the absence of dynamic variation (the buzzing sound colour makes it impossible to vary the loudness), the specific flowing rhythm and the softened voice tone induces a very meditative, introspective mood, distinct from other Amhara musical expressions (Weisser, 2012).

16.2.1 Bagana Playing and Notation

The bagana can be played by plucking the strings with the left hand fingers or with a plectrum (*girf*). The latter technique has nowadays almost disappeared, even though iconographic evidences (see Fig. 16.2: left) suggest that it was frequently in use. The bagana is a monophonic instrument; even when played with the plectrum, the musician selects the pitches by allowing to vibrate freely only the strings that are supposed to sound, and by blocking with the fingers the ones that are not supposed to sound. When plucked with the fingers, the strings are also played individually, even though, as the strings are left open and therefore sound until natural extinction, some superposition might be created when the plucking of several strings is fast.

The plucking technique of bagana consists in placing the left hand behind the strings, and performing a movement from a non-sounding string (a rest string) to the next one (a playing string), to pluck it with the flesh of the finger and then back to the rest string. Traditionally, the playing of the bagana was taught privately, by a master player to a pupil. The pupil placed his hand on the master's and understood that way how to proceed. The musical part of bagana repertoire was not written down: only lyrics were sometimes notated. However, the development of musical teaching inspired from Western music schools led to the opening of a bagana class at the Yared School of Music (Addis Ababa). The professor Alemu Aga, one of the most renowned and acclaimed bagana players of the country, who taught this class from

Table 16.1 Fingering of the bagana, with finger numbers assigned to string numbers

string	1	2	3	4	5	6	7	8	9	10
finger	1	r	$2'$	2	r	3	r	4	r	5

1972 to 1980, developed a new method in order to formalize the learning process, in cooperation with the professors of other instruments. He assigned a number for each finger, each sounding string and each pitch (see Table 16.1 and Fig. 16.3).

In Table 16.1, "r" (for "rest") indicates a string that is not played, but rather is used as a rest for the finger after it plucks the string immediately next to it. Strings 3 and 4 are both played by finger number 2 (string 3 being therefore notated as finger $2'$), otherwise the assignment of finger number to string number is fixed.

Only six of the ten strings are playing strings, tuned according to a pentatonic scale. The two scales (*keniet*, pl. *kenietotch*) used for bagana performance are usually either anhemitonic (with all intervals larger than a tempered semitone), or hemitonic (containing one or more semitones). The anhemitonic scale is called *tezeta* and the hemitonic one is often named *anchihoye* (see Table 16.2). As in the rest of Amhara music, both secular and sacred, the scales are built on untempered intervals and the absolute pitch is irrelevant. Players usually adjust their pitch range to their taste and voice range, transposing the scale to the desired voice register.

To illustrate the notation in Table 16.2, for example, in the tezeta scale the ascending pentatonic scale (closest tempered degrees being C, D, E/F, G, A) would be notated by the sequence $[2,3,1,5,4]$. The scale degree notation is also useful for considering the intervallic relation between two strings, as will be used in Sect. 16.4. Figure 16.3 shows the placement of the left hand on the 10 bagana strings, along with the information from Table 16.2: the finger numbering used, and the notes played by the fingers (in the tezeta scale). Figure 16.4 shows an example of a fragment of a bagana song, encoded as a sequence of finger numbers, corresponding to the fingering of the song.

Table 16.2 Tuning of the bagana, in two different scales, and the nearest Western tempered note (with octave number) corresponding to the degrees of the scales. String 1 in the tezeta scale can be tuned to either E2 or F2

finger	1	$2'$ and 2	3	4	5
string	1	3 and 4	6	8	10
scale tezeta	E2 or F2	C2	D2	A2	G2
scale anchihoye	F2	C2	D♭2	A2	G♭2
scale degree	$\widehat{3}$	$\widehat{1}$	$\widehat{2}$	$\widehat{5}$	$\widehat{4}$

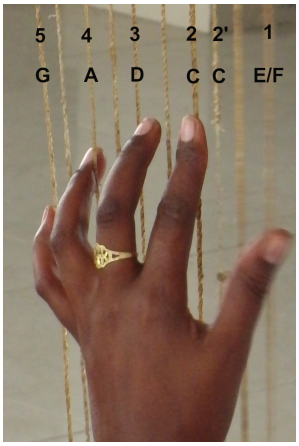

Fig. 16.3 Placement of left hand on the strings of the bagana. Photo: Stéphanie Weisser, Addis Ababa, October 2012

16.2.2 Known Motifs

In the context of formal teaching, Alemu Aga wrote down with this numbered notation all of the melodic material of the songs he performed. He also designed specific exercises: based on an analysis of his personal repertoire, he developed several motifs (two to seven notes), to be used in an early stage of the learning process (see Tables 16.3 and 16.4). The student, after learning how to hold the instrument properly and placing the left hand in a correct position, is expected to practise these series of motifs regularly until moving on to learning a real song. To this day, Alemu Aga still uses these motifs in his teaching, considering that they familiarize the student with the playing technique, the numbered notational system, the pitch of each string, and the unique buzzing sound colour of the instrument.

In the teaching process, Alemu Aga requests the student to practise each exercise three times before moving to the next one. When the motif is short (such as the first and second of the rare motifs in Table 16.3, and the second and third of the frequent motifs in Table 16.4), repetition is usually performed without a break, which leads the student to practice both the motif and its retrograde. For example, the pattern $[1,4,1,4,1,4]$ (repeating the first rare motif three times) effectively comprises the pattern $[1,4]$ and its retrograde $[4,1]$. When the exercise is long (such as the third and fourth of the rare motifs, and the first of the frequent motifs), a pause is usually taken

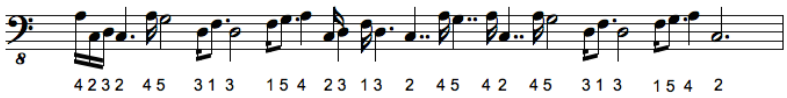

Fig. 16.4 A fragment encoded in score and finger notation, from the beginning of the song *Abatachen Hoy* ("Our Father"), one of the most important bagana songs, as performed by Alemu Aga (voice transcription and lyrics not shown). Transcribed by Weisser (2006)

Table 16.3 The four rare motifs provided by Alemu Aga, from Weisser (2005, page 50)

	Motifs in numeric notation
First exercise	$[1,4]$
Second exercise	$[1,2]$
Third exercise	$[2,3,1,5,4]$
Fourth exercise	$[4,5,1,3,2]$

Table 16.4 The three frequent motifs provided by Alemu Aga, from Weisser (2005, page 50)

	Motifs in numeric notation
First exercise	$[4,5,4,5,4,5,1]$
Second exercise	$[4,2]$
Third exercise	$[3,1]$

after the final note, before starting the motif all over again. Therefore, it cannot be said that the bigrams $[4,2]$ and $[2,4]$ (which are the joining bigrams formed by the repetition of the third and fourth rare motifs) are rare, nor that the motif $[1,4]$ (which is the joining bigram formed by the repetition of the first frequent motif) is frequent. In fact, the opposite is true ($[1,4]$ is a rare motif: see Table 16.3) and this will also be confirmed by the pattern discovery results of Sect. 16.4.

16.2.2.1 Rare Motifs

Table 16.3 shows four motifs that correspond, according to the bagana master Alemu Aga, to motifs that are rarely encountered in his real bagana songs and are used during practice to strengthen the fingers with unusual finger configurations (Weisser, 2005). Referring to Fig. 16.4, it can be seen that the motifs of Table 16.3 are absent from the fragment. The first two motifs are bigrams, and in Sect. 16.4 it will be explored whether these two rare motifs can be discovered from corpus analysis. The third and fourth motifs correspond to longer pentagrams that form ascending and descending pentatonic scales and are also used for didactic purposes. Since most pentagram patterns will be rare in a small corpus, an additional question that will be explored in Sect. 16.4 is whether these two pentagram motifs are *surprisingly* rare.

16.2.2.2 Frequent Motifs

Table 16.4 shows three motifs that correspond, according to the bagana master Alemu Aga, to motifs that are frequently encountered in his real bagana songs. The first exercise works the independence of the little finger regarding the ring finger and increases the strength of the little finger. The ring finger and the little finger are particularly exercised because they pluck the tautest strings of the instrument. The final part $[5,1]$ of the first exercise also directly outlines the particular difficulty of the thumb, which plucks the string in the opposite direction compared to the other fingers. The second exercise focuses on mastering extreme notes of the ambitus of the bagana. This motif is very often used, particularly at the beginning and at the end of musical phrases (for example, see the final notes of the fragment of Fig. 16.4).

Finally, the third exercise of this category outlines the difficulty of plucking a string with the middle finger without moving the ring finger.

16.2.3 Bagana Corpus

The analysed corpus comprises 29 melodies of bagana songs and 8 derderotch (instrumental preludes) performed by seven players (five men, two women). These 37 pieces were recorded and transcribed by Weisser (2005) between 2002 and 2005 in Ethiopia (except for two of them recorded in Washington, DC).

It is important to note that in bagana performances, melodic variations are frequent, and the performance is rarely completely fixed. However, these modifications are not structural, as they do not undermine the identity of the song. Some musicians introduce these changes in a different way each time they are playing the song. In order to limit the impact of these variations on the results, non-structural variations were not included, nor were additional and optional ornamentations.

After removal of non-structural variations, a total of $N = 1906$ events (finger numbers) are encoded within the 37 pieces (events per song: $\mu = 51$, $\sigma = 30$, min $=$ 13, max $= 121$).

16.3 Pattern Discovery Method

In this work we apply data mining to discover distinctive patterns in the bagana corpus. Here a pattern P of length ℓ is a contiguous sequence of finger numbers, notated as $[e_1, \ldots, e_\ell]$. The number of occurrences of a pattern P in the corpus is given by c_P (see Table 16.5).

In pattern discovery in music, the counting of pattern occurrences can be done in two ways (Conklin, 2010b): either by considering *piece count* (the number of pieces containing the pattern one or more times, i.e., analogous to the standard definition of pattern support in sequential pattern mining) or by considering *total count* (the total number of positions matched by the pattern, also counting multiple occurrences within the same piece). Total count is used whenever a single piece of music is the target of analysis. For the bagana, even though several pieces are available, total count is used, because we consider that a pattern is frequent (or rare) if it is frequently (or rarely) encountered within any succession of events. Therefore in this study c_P is the total number of events which initiate a pattern P.

Table 16.5 Glossary of notation and terminology used in this chapter

notation / terminology	meaning
pattern	a sequence of finger numbers
$P = [e_1, \ldots, e_\ell]$	pattern P of length ℓ
motif	a known frequent or rare pattern for the bagana
positive pattern	over-represented in a corpus
antipattern	under-represented in a corpus
c_P	total count of pattern P
c_e	total count of event e
N	total number of events in corpus
X	random variable modelling the total count c_P of pattern P
t_P	maximum possible total count of pattern P
b_P	analytic background probability of pattern P
$\mathbb{E}$	expected total count of pattern
$\mathbb{B}$	binomial probability density function
$\mathbb{B}_\leq$	binomial cumulative distribution function
α	statistical significance level

16.3.1 Pattern Statistics

In this section two subtypes of pattern are introduced. A *positive pattern* is a pattern which is frequent in a corpus; an *antipattern* is a sequence that is rare, or even absent, in a corpus. For data mining, these definitions though intuitive are not operational. For positive patterns, very short patterns will tend to be highly frequent but usually are not meaningful because they may occur with high frequency in any corpus. For antipatterns, almost any sequence of events is an antipattern, that is, most possible event sequences will never occur in a corpus, and most are not interesting because it is expected that their total count is zero. Therefore we want to know which are the *significant* positive patterns and antipatterns: those that are *surprisingly* frequent or rare in a corpus.

Positive patterns and antipatterns can be evaluated by their *p-value*, which gives the probability of finding at least (for positive patterns) or at most (for antipatterns) the observed number of occurrences in the corpus. The binomial distribution, which can be used to compute the probability of obtaining an observed number of occurrences in a given number of sequence positions, is a standard model for assessing discovered motifs in bioinformatics (van Helden et al., 1998) and will be used here to compute pattern p-values as described below.

In studies where there is a set of pieces available to contrast with the corpus, it is possible to compute the *empirical background probability* of a pattern. This was the method used for ranking patterns in inter-opus pattern discovery studies of Cretan folk songs (Conklin and Anagnostopoulou, 2011) and antipatterns in Basque folk tunes (Conklin, 2013). For the bagana there is no natural set of pieces to contrast with the corpus, and the background probability of a pattern $P = [e_1, \ldots, e_\ell]$ must be estimated, for example using a zero-order model of the corpus:

$$b_P = \prod_{i=1}^{\ell} (c_{e_i}/N) \;, \tag{16.1}$$

where c_{e_i} is the total count of event e_i, and N is the total number of events in the corpus. This *analytic background probability* b_P therefore estimates the probability of finding the pattern in ℓ contiguous events. In this study, a zero-order model of the corpus is used: a higher-order analytic model would not be able to detect bigram patterns because the expected total count of a bigram pattern would be equivalent to its actual count.

To define the p-value of a pattern $P = [e_1, \ldots, e_\ell]$, we define the random variable X that describes its total count c_P. This is modelled by the binomial distribution (see Appendix A for definitions):

$$\mathbb{P}\big(X = c_P\big) = \mathbb{B}\big(c_P; t_P, b_P\big) \;, \tag{16.2}$$

where $t_P = \lfloor N/\ell \rfloor$ approximates the maximum number of positions that can be possibly matched by the pattern (this is a lower bound, as patterns with self-overlap have a higher maximum) and b_P is the background probability of the pattern.

Letting $\mathbb{B}_{\leq}$ be the binomial cumulative distribution function (see Appendix A for definitions), the p-value of P as an antipattern is the probability of finding c_P or fewer occurrences of the pattern in the corpus:

$$\mathbb{P}\big(X \leq c_P\big) = \mathbb{B}_{\leq}\big(c_P; t_P, b_P\big) \;, \tag{16.3}$$

and the p-value of P as a positive pattern is the probability of finding c_P or more occurrences of the pattern in the corpus:

$$\mathbb{P}\big(X \geq c_P\big) = 1 - \mathbb{B}_{\leq}\big(c_P - 1; t_P, b_P\big) \;. \tag{16.4}$$

Low p-values according to (16.3) or (16.4) indicate patterns that are statistically surprising and therefore potentially interesting.

16.3.2 Pattern Discovery Algorithm

The pattern discovery task can be stated as: given a corpus, find all patterns having a p-value ((16.3) and (16.4)) below the specified significance level α. Furthermore,

for presentation we consider only those significant patterns that are *minimal*, that is, those that do not contain any other significant pattern (Conklin, 2013; Ji et al., 2007).

The discovery of minimal significant patterns can be efficiently solved by a refinement search over the space of possible patterns (Conklin, 2010a, 2013), using an algorithm similar to general approaches for sequential pattern mining (Ayres et al., 2002). The refinement search is also similar to the STUCCO (Bay and Pazzani, 2001) method for contrast set mining, here with a method for under-represented pattern pruning. A depth-first search (see Fig. 16.5) starts at the most general (empty) pattern. The search space is structured lexicographically so that no pattern is visited twice. If a pattern P at a particular node of the search tree is not significant (either as a positive pattern or an antipattern: see (16.3) and (16.4)), it is *specialized* to a new pattern P' by extending it on the right hand side by every possible event (finger number) and updating the pattern counts $c_{P'}$, and computing $t_{P'}$, $b_{P'}$, and both p-values ((16.3) and (16.4)). All specializations of the pattern are then added to the search queue. The search continues at a node if the pattern is significant neither as a pattern nor as an antipattern. Significant patterns are added to a solution list, with a filtering against other solutions found so far to ensure that the final pattern set contains only minimal patterns.

The runtime of the algorithm is largely determined by the specified significance level α, because with low α the search space must be more deeply explored before a significant positive pattern is reached. For positive patterns, the p-value (defined in (16.3)) rapidly decreases with pattern specialization, so unless α is very low the search for positive patterns will tend to terminate early. For antipatterns the situation is reversed, with the p-value tending to increase through a search path.

Nevertheless, for antipatterns it is possible to compute the minimal antipattern p-value (16.3) achievable on a search path. This can lead to the pruning of entire

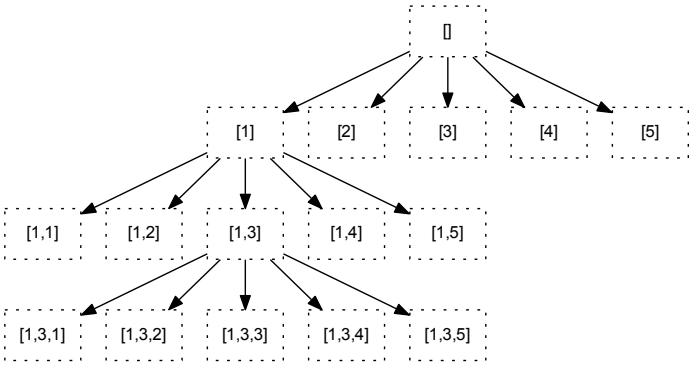

Fig. 16.5 A refinement search space of bagana patterns, through the pattern $[1,3]$ down to its specializations

search paths that cannot possibly visit an antipattern meeting the significance level of α:

Theorem 16.1. *For any pattern P, if* $\mathbb{B}\big(0;t_P,b_P\big) > \alpha$*, then* $\mathbb{B}_{\leq}\big(c_{P'};t_{P'},b_{P'}\big) > \alpha$ *for any specialization* P' *of P.*

Proof. Consider a pattern $P = [e_1,\ldots,e_\ell]$ and a specialization $P' = [e_1,\ldots,e_{\ell'}]$. Since $\ell' > \ell$, it follows that $b_P \geq b_{P'}$ (16.1) and that $t_P \geq t_{P'}$. This implies that $(1-b_P)^{t_P} \leq (1-b_{P'})^{t_{P'}}$. Therefore, $\mathbb{B}\big(0;t_P,b_P\big) \leq \mathbb{B}\big(0;t_{P'},b_{P'}\big)$. Following the definition of the cumulative distribution $\mathbb{B}_{\leq}$ (16.6), $\mathbb{B}\big(0;t_{P'},b_{P'}\big) \leq \mathbb{B}_{\leq}\big(c_{P'};t_{P'},b_{P'}\big)$. Therefore $\mathbb{B}\big(0;t_P,b_P\big) \leq \mathbb{B}_{\leq}\big(c_{P'};t_{P'},b_{P'}\big)$, so if $\mathbb{B}\big(0;t_P,b_P\big) > \alpha$, then $\mathbb{B}_{\leq}\big(c_{P'};t_{P'},b_{P'}\big) > \alpha$. □

The impact of this new theorem is that if the pattern P at a search node is being tested only as an antipattern, and $\mathbb{B}\big(0;t_P,b_P\big) > \alpha$, the search does not need to continue at that search node. The theorem also provides an important way to use the algorithm above for efficient *unword* discovery, i.e., to find significant antipatterns where $c_P = 0$.

16.4 Results and Discussion

The pattern discovery method described in Sect. 16.3 was applied to the bagana corpus to find all minimal antipatterns and positive patterns at the significance level of $\alpha = 0.01$. In this section, first the discovered antipatterns, then the discovered positive patterns, are presented and discussed.

16.4.1 Discovered Antipatterns

Table 16.6 presents the antipatterns revealed by the discovery method. In the columns are the pattern P, the closest tempered interval (in the tezeta scale) formed by the components of the pattern, the pattern counts (both total count and piece count), the expected total count of P given by $\mathbb{E}(X) = b_P \times t_P$, and finally the p-value according to (16.3).

The method revealed exactly ten significant antipatterns: five antipatterns and their retrogrades. Interestingly, all are bigrams and all come in retrograde pairs. The two antipatterns $[4,1]$ and $[2,1]$ presented in the top part of Table 16.6 are the *most significant* antipatterns discovered, and correspond to the retrogrades of two of the didactic rare motifs (Table 16.3). As mentioned earlier, this discovery is valid as, for example, the first motif $[1,4]$ repeated three times will also play the motif $[4,1]$. It is therefore remarkable that both retrograde forms, not only the form listed in Table 16.3, are discovered.

The second column of Table 16.6 presents the closest tempered interval (in the tezeta scale) formed by the pattern. Note that antipatterns involving finger 1 can give

Table 16.6 Bagana antipatterns discovered (top and middle) at $\alpha = 0.01$. Intervals are measured in the tezeta scale. Top: corresponding to rare bigram motifs of Table 16.3; middle: novel rare patterns; bottom: the two pentagram motifs of Table 16.3. Partial horizontal lines separate retrograde pairs. Numbers in brackets indicate the number of distinct pieces containing the pattern

P	closest tempered interval	c_P	$\mathbb{E}(X)$	$\mathbb{P}(X \leq c_P)$
$[1,4]$	P4/M3	21 (14)	48	7.5e-06
$[4,1]$		2 (2)	48	6.3e-19
$[2,1]$	M3/P4	6 (5)	50	1.8e-15
$[1,2]$		13 (7)	50	2.6e-10
$[3,4]$	P5	3 (3)	30	4.0e-10
$[4,3]$		2 (2)	30	3.8e-11
$[2,5]$	P5	16 (9)	38	2.7e-05
$[5,2]$		17 (10)	38	6.6e-05
$[3,5]$	P4	5 (4)	26	5.7e-07
$[5,3]$		11 (8)	26	0.00077
$[2,3,1,5,4]$	ascending scale	8 (7)	0.11	1.0
$[4,5,1,3,2]$	descending scale	4 (3)	0.11	1.0

rise to two alternatives, depending on the tuning of string 1 to either an E or an F (see Table 16.2). Interestingly, all of the discovered antipatterns form a melodic interval of a major third or greater (intervals of P4, M3, and P5).

In addition to the two bigram motifs in Table 16.3, the three additional antipatterns $[3,4]$, $[2,5]$, and $[3,5]$ (with their retrogrades) are found by the method. It is worth noticing specifically the rarity of the interval of fifths (perfect in tezeta, diminished and augmented in anchihoye) outlined by antipatterns $[3,4]$ and $[2,5]$. According to authoritative writings in ethnomusicology (Arom, 1997), the perfect fifth and the cycle of fifths play a founding role in anhemitonic pentatonic scales such as tezeta. The rarity of perfect fifths (P5) in the songs is significant, and it can be speculated that this interval is a mental reference that is never (and does not necessarily need to be) performed. Similarly, the antipattern $[3,5]$, a perfect fourth (P4), is the inversion of a perfect fifth. The rarity of the fourth can be related to the rarity of the fifth, as intervals and their inversions are usually connected in several musical cultures, including Western art music.

For completeness with the results of Weisser (2005), at the bottom of Table 16.6 are the two pentagram motifs from Table 16.3. As expected, these motifs are not frequent, but it is notable that they are not significant as antipatterns according to their p-value (therefore they are not reported by the pattern discovery method). In fact, they occur in the corpus more than expected according to their background probability.

16.4.2 Discovered Positive Patterns

The method revealed fourteen significant positive patterns (Table 16.7): five patterns and their retrogrades (reversal), and four unison patterns (repetitions of the same finger number). Most of the significant patterns are bigrams, as with the antipatterns in Table 16.6. The unison patterns are not presented further: according to bagana musicians, the exact number of a repeated pitch is not important and may depend on other factors than the structure of music.

The three patterns $[3,1]$, $[2,4]$, $[5,4]$ (with their retrogrades, which are also significant) cover the frequent motifs in Table 16.4. As with the rare motifs, the discovery of the retrograde is valid, as the frequent motifs are also practised three times in a row. Regarding the discovered pattern $[1,5]$, which is not listed in Table 16.4, it is the retrograde of $[5,1]$ (which is the end of the first frequent motif of Table 16.7). It is worth noticing that the pattern $[1,5]$, though appearing in Alemu Aga's third rare motif (see Table 16.3), is only a small part of this motif: $[1,5]$ is therefore not practised as much as the other retrogrades. However, the fact that the discovered pattern $[1,5]$ is found among the rare motifs can be explained: it is mostly in songs studied in a more advanced phase in the learning process. Therefore, the practice of the $[1,5]$ motif in a familiarization phase is not really needed. Its discovery as a pattern is consistent with all discovered bigrams being in retrograde pairs.

In terms of melodic intervals, the discovered patterns represent conjunct melodic motion (one scale step), except for the pattern $[2,4]$, corresponding in the tezeta scale to a major sixth and the largest interval playable on the instrument. This pattern is also considered idiomatic of bagana playing. Interestingly, the two pentagram motifs of Table 16.3, presented according to their rarity in the songs of Alemu Aga, are composed entirely of frequent significant bigram motifs.

Table 16.7 Bagana positive patterns discovered at $\alpha = 0.01$. Top: frequent bigram motifs of Table 16.4; bottom: two novel discovered patterns. Partial horizontal lines separate retrograde pairs. Numbers in brackets indicate the number of distinct pieces containing the pattern

P	closest tempered interval	c_P	$\mathbb{E}(X)$	$\mathbb{P}(X \geq c_P)$
$[3,1]$	M2/m3	195 (37)	34	1e-87
$[1,3]$		148 (35)	34	4.6e-51
$[2,4]$	M6	154 (35)	44	9.4e-41
$[4,2]$		200 (37)	44	1.3e-71
$[5,4]$	M2	127 (31)	37	3.7e-33
$[4,5]$		112 (31)	37	8e-25
$[1,5]$	m3/M2	160 (35)	41	6.4e-48
$[5,1]$		139 (32)	41	4.9e-35
$[2,3]$	M2	104 (35)	31	3.7e-26
$[3,2]$		67 (28)	31	7.7e-09

It can be hypothesized that the favouring of patterns forming a conjunct melodic motion is linked with the singing voice. Indeed, as the lyrics are of great importance and mostly syllabic (each syllable is sung on one note), such a conjunct motion is easier for the musician to sing as well as easier for the listener to understand. It should be noted that the singing voice accompanying the bagana is not limited to a total range of a sixth: as an interval of an octave is not considered significant, the pattern $[4, 2]$ can be also performed as conjunct, when the pitch of string 4 (A) is sung an octave lower than usual, meaning lower than the pitch of string 2 (C).

16.5 Conclusions

An inductive approach to computational music analysis holds great potential for revealing both *intra-opus* motifs from a single piece, or *inter-opus* motifs from a corpus of several pieces. In cases where motifs of functional significance are known, pattern discovery methods can be validated against prior knowledge, and new discoveries may gain potential musicological significance.

This chapter has developed an inter-opus pattern discovery method and applied it to the discovery of over- and under-represented patterns in bagana songs. From a small corpus of bagana songs, the method was able to find the rare and frequent motifs used by Alemu Aga for bagana teaching. The validation of pattern discovery on a known motif set is invaluable as it lends additional interest to novel motifs discovered, and even to future work with other corpora.

A zero-order background model for pattern discovery provided excellent results for pattern discovery in the bagana corpus, with few type I errors (discovered minimal patterns that are not bigram motifs) and no type II errors (bigram motifs that are not discovered as minimal patterns). As mentioned above, higher-order background models would not be able to detect the types of short motifs that are known for the bagana. The selection of a background model for pattern discovery, necessary in music analysis tasks where no explicit anticorpus is available, is always an issue in statistical hypothesis testing, being a tradeoff between tolerance of type I and type II errors. In the field of musicology, Huron (1999) refers to the reluctance to discard patterns as "theory-conserving skepticism", arguing that the application of statistical methods to musicology should have high statistical power to avoid overlooking any potentially interesting pattern.

Antipattern discovery in music, the process of discovering patterns that are surprisingly rare in a group of pieces, was introduced in the context of Basque folk tunes (Conklin, 2013), where each annotated tune genre is contrasted against remaining genres. This chapter shows that antipattern discovery can be carried out without an explicit set of contrasting pieces by setting up a background statistical model that produces expected counts in the corpus. Large deviations have low probabilities according to the binomial distribution, and are reported as significantly rare patterns.

From a musicological point of view, the analysis we present here can be characterized by its *emic* basis, based on a musician *within* the culture. Such ground provides

the analyst with a very important clue regarding the way a musical piece is built and conceptualized. Influenced by linguistics and more specifically by phonology, (ethno)musicologists often search for "minimal units" in a musical repertoire, similar to phonemes in a language. In the case of the bagana, the use of such motifs *by the musicians themselves* is a precious indication towards the idea that the minimal units of bagana songs are not single pitches, but rather pairs (or triplets) of sounds.

Such formal principles made the computational analysis of the presence of motifs pertinent in this specific context. The analysis of non-Western, non-written (or partially non-written) repertoires is a complex task, mostly when it comes to the validation of hypotheses constructed by the analyst—often an outsider to the cultural and musical system being studied. Several researchers, including Arom (1997), developed methods allowing a kind of generation of the repertoire according to the hypothesis tested by the investigator (Fernando, 2004). Indeed, with such a method, the musician can therefore validate a result, instead of a concept. Pattern and antipattern discovery, with its inductive approach, paves the way for generating music according to structural and formal rules. Such an approach will be invaluable for the analysis of a repertoire, especially in the field of ethnomusicology.

Acknowledgements This research is supported by the project Lrn2Cre8 which is funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859. Special thanks to Kerstin Neubarth for valuable comments on the manuscript.

A Binomial Distribution: Definitions and Notation

Binomial probability density function (probability of exactly k successes in n independent trials, each having probability p of success):

$$\mathbb{B}\big(k;n,p\big) = \binom{n}{k} p^k (1-p)^{n-k} \,. \tag{16.5}$$

Binomial cumulative distribution function (probability of k or fewer successes in n independent trials, each having probability p of success):

$$\mathbb{B}_{\leq}\big(k;n,p\big) = \sum_{i=0}^{k} \mathbb{B}\big(i;n,p\big) \,. \tag{16.6}$$

References

Adamo, J.-M. (2001). *Data Mining for Association Rules and Sequential Patterns*. Springer.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan.

Arom, S. (1997). Le "syndrome" du pentatonisme africain. *Musicae Scientiae*, 1(2):139–163.

Artamonova, I., Frishman, G., and Frishman, D. (2007). Applying negative rule mining to improve genome annotation. *BMC Bioinformatics*, 8:261.

Ayres, J., Gehrke, J., Yiu, T., and Flannick, J. (2002). Sequential pattern mining using a bitmap representation. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 429–435, Edmonton, Canada.

Bay, S. and Pazzani, M. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246.

Collins, T., Böck, S., Krebs, F., and Widmer, G. (2014). Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio. In *Proceedings of the Audio Engineering Society's 53rd Conference on Semantic Audio*, London, UK.

Conklin, D. (2009). Melody classification using patterns. In *International Workshop on Machine Learning and Music*, pages 37–41, Bled, Slovenia.

Conklin, D. (2010a). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554.

Conklin, D. (2010b). Distinctive patterns in the first movement of Brahms' String Quartet in C Minor. *Journal of Mathematics and Music*, 4(2):85–92.

Conklin, D. (2013). Antipattern discovery in folk tunes. *Journal of New Music Research*, 42(2):161–169.

Conklin, D. and Anagnostopoulou, C. (2001). Representation and discovery of multiple viewpoint patterns. In *Proceedings of the International Computer Music Conference*, pages 479–485, Havana, Cuba.

Conklin, D. and Anagnostopoulou, C. (2011). Comparative pattern analysis of Cretan folk songs. *Journal of New Music Research*, 40(2):119–125.

Deng, K. and Zaïane, O. R. (2009). Contrasting sequence groups by emerging sequences. In Gama, J., Santos Costa, V., Jorge, A., and Brazdil, P., editors, *Discovery Science*, volume 5808 of *Lecture Notes in Artificial Intelligence*, pages 377–384. Springer.

Fernando, N. (2004). Expérimenter en ethnomusicologie. *L'Homme*, 171-172:284–302.

Forte, A. (1983). Motivic design and structural levels in the first movement of Brahms's String Quartet in C minor. *The Musical Quarterly*, 69(4):471–502.

Haglin, D. J. and Manning, A. M. (2007). On minimal infrequent itemset mining. In *Proceedings of the 2007 International Conference on Data Mining*, pages 141–147, Las Vegas, Nevada.

Herold, J., Kurtz, S., and Giegerich, R. (2008). Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, 9:167.

Hirao, M., Hoshino, H., Shinohara, A., Takeda, M., and Arikawa, S. (2000). A practical algorithm to find the best subsequence patterns. In Arikawa, S. and Morishita,

S., editors, *Discovery Science*, volume 1967 of *Lecture Notes in Computer Science*, pages 141–154. Springer.

Huron, D. (1999). The new empiricism: Systematic musicology in a postmodern age. Lecture 3 from the 1999 Ernest Bloch Lectures. http://musiccog.ohio-state.edu/Music220/Bloch.lectures/3.Methodology.html. Accessed Mar 3, 2015.

Huron, D. (2001). What is a musical feature? Forte's analysis of Brahms's Opus 51, No. 1, revisited. *Music Theory Online*, 7(4).

Ji, X., Bailey, J., and Dong, G. (2007). Mining minimal distinguishing subsequence patterns with gap constraints. *Knowledge and Information Systems*, 11(3):259–296.

Lartillot, O. (2004). A musical pattern discovery system founded on a modeling of listening strategies. *Computer Music Journal*, 28(3):53–67.

Lin, C.-R., Liu, N.-H., Wu, Y.-H., and Chen, A. (2004). Music classification using significant repeating patterns. In Lee, Y., Li, J., Whang, K.-Y., and Lee, D., editors, *Database Systems for Advanced Applications*, volume 2973 of *Lecture Notes in Computer Science*, pages 506–518. Springer.

Meredith, D., Lemström, K., and Wiggins, G. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345.

Mooney, C. H. and Roddick, J. F. (2013). Sequential pattern mining – approaches and algorithms. *ACM Computing Surveys*, 45(2):19:1–19:39.

Sawada, T. and Satoh, K. (2000). Composer classification based on patterns of short note sequences. In *Proceedings of the AAAI-2000 Workshop on AI and Music*, pages 24–27, Austin, Texas.

Shan, M.-K. and Kuo, F.-F. (2003). Music style mining and classification by melody. *IEICE Transactions on Information and Systems*, E88D(3):655–659.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281(5):827–842.

Wang, J., Zhang, Y., Zhou, L., Karypis, G., and Aggarwal, C. C. (2009). CONTOUR: an efficient algorithm for discovering discriminating subsequences. *Data Mining and Knowledge Discovery*, 18(1):1–29.

Webb, G. I. (2007). Discovering significant patterns. *Machine Learning*, 68(1):1–33.

Weisser, S. (2005). *Etude ethnomusicologique du bagana, lyre d'Ethiopie*. PhD thesis, Université libre de Bruxelles.

Weisser, S. (2006). Transcrire pour vérifier: le rythme des chants de bagana d'Éthiopie. *Musurgia*, XIII(2):51–61.

Weisser, S. (2012). Music and Emotion. The Ethiopian Lyre Bagana. *Musicae Scientiae*, 16(1):3–18.

Wu, X., Zhang, C., and Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, 22(3):381–405.