行政院國家科學委員會專題研究計畫　期中進度報告

子計畫一:視訊的智慧型高階處理(1/3)

計畫類別：整合型計畫
計畫編號：NSC91-2219-E-002-043-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立臺灣大學電機工程學系暨研究所

計畫主持人：貝蘇章

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 15 日

智慧型音視訊和傳輸技術及多媒體應用-子計畫一：視訊的智慧型高階處理(I)

Intelligent High level Video Processing (I)

計畫編號：NSC-91-2219-E-002-043

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：貝蘇章　台灣大學電機系教授

摘要

我們發展一些影像二元化的有效方法，進而延伸靜態影像到動態視訊方面，有效率的二階視訊系統被開發，可以應用到在低頻寬無線環境下的影像視訊傳輸系統。

關鍵字: 影像二元化、二階視訊

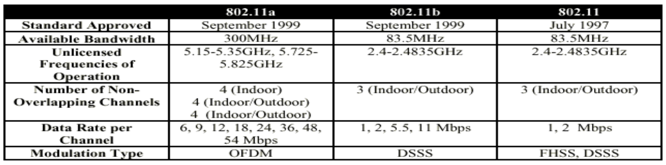
Abstract

We have developed some global binarization methods and proposed two locally adaptive methods. In this report, we will extend the application from static images to dynamic videos. A bi-level video system will be introduced and constructed in this report. It can be used in low bandwidth wireless video transmission system.

Keyword: Image binarization、bi-level video

Introduction

The rapid development of wired and wireless networks tremendously facilitates communications between people. However, most of the current wireless networks still work in low bandwidths. For example, GSM system has a bandwidth limitation of 9.6 Kbps, and after applying GPRS service on it, theoretically the transmission speed can be up to 171.2 Kbps, but in reality only 40-50 Kbps [37]. This bandwidth is sufficient given that only voice or audio information is transmitted. However, other data types like video require much more bandwidth and still cannot be accessed through GSM mobile devices. Although wireless LAN 802.11 has offered higher transmission speed ranges from 1-54 Mbps, the geographic transmission range for each access point has a limit of a few hundred meters. Besides, interference and security issues remain great difficulties for WLAN communication.



	802.11a	802.11b	802.11
Standard Approved	September 1999	September 1999	July 1997
Available Bandwidth	300MHz	83.5MHz	83.5MHz
Unlicensed Frequencies of Operation	5.15-5.35GHz, 5.725-5.825GHz	2.4-2.4835GHz	2.4-2.4835GHz
Number of Non-Overlapping Channels	4 (Indoor) 4 (Indoor/Outdoor) 4 (Indoor/Outdoor)	3 (Indoor/Outdoor)	3 (Indoor/Outdoor)
Data Rate per Channel	6, 9, 12, 18, 24, 36, 48, 54 Mbps	1, 2, 5.5, 11 Mbps	1, 2 Mbps
Modulation Type	OFDM	DSSS	FHSS, DSSS

Table 1 IEEE 802.11 standards

While video on demand (VOD) or video conference are achievable in wired communication, the counterparts in wireless communication are void because of the low bandwidth, and mobile devices also suffer from weak computational power, short battery lifetime and limited display capability. Although conventional video processing schemes such as MPEG 1/2/4 [38] and H.261/263 [39 40] can also code video for very low bitrates, the resultant images usually look like a collection of color blocks and the motion in the scene becomes discontinuous. The block effect of these methods originates from the common architecture of discrete cosine transform (DCT) based coding. The image is first divided into pixel blocks and then DCT is employed to transform the image into frequency domain. These DCT coefficients are then quantized using quantization table and compressed using entropy encoder. As a result, the low spatial frequency components representing the average values of each block possess high priority. Thus, if DCT-based compression methods work at limited bandwidths, the basic color of blocks will be kept in preference. This will cause block effect and blur the original object contours in the images. If we make use of binarization techniques, bit rate of video sequences can be effectively reduced while retaining sufficient information. Take video conference as an example, the facial expressions are essential information in conversation, thus, outlines of the face, eyes, eyebrows, and mouth should have more importance than block averages of facial regions. In binarized images, contour information is better preserved than DCT-based methods under low bandwidth constraint. While low bitrate feature is favorable under the low bandwidth limitation of wireless communication, in the broadband wired communication, binarized video can also find its applications in various fields such as video indexing, video preview, surveillance system…etc [41].

In this report, we will focus on the binarization of video sequences on the encoder end. The principles for this bi-level video system should be *low complexity*, *low computation demands*, *low bit rates*, and *sufficient information preservation*. In addition to the application of proposed binarization, we devices two more functions that can further reduce the bit rate and emphasize skin tone regions, respectively.

,

Bi-Level Video System Flowchart

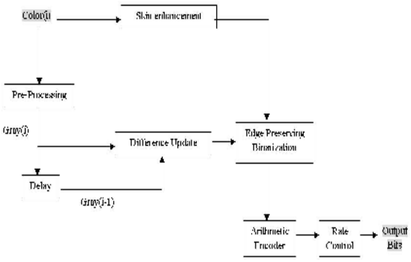


Figure 1. Bi-level video system flowchart

In the block diagram, we can see some of the previously introduced functions. Such as pre-processing function that transforms color input images into gray level ones, this function is extracted out from edge preserving binarization to better meet other functions' requirement. Edge preserving binarization is as described in section 3, and the setting for edge detection is Laplacian second order edge detection. Note that an extra input from skin enhancement block is present at the input of edge preserving binarization, this skin tone region information will be incorporated into the binarization block. Skin enhancement and difference update block are devised for the video system and will be discussed in depth in this chapter. After the binarization of video, it should be further coded for compression and transmission, arithmetic encoder and rate control blocks are then utilized. However, while we focus on the binarization of video sequences, these two blocks will not be included in this chapter.

Skin Detection

Skin tone regions in the videos include faces, hands, and bodies of human beings, and they provide essential information like human locations, emotional expressions shown on the faces, gestures of the hands…etc. In other words, regions in the video with humans in them are of more perceptual importance and worth additional attention and enhancement. A successful skin detection method should be robust to facial shape, size, skin color, orientation, motion, and lighting condition. We will adopt skin detection using color space segmentation technique in HSV color space for this function block. A skin tone area is first defined in HSV color space, and the image is examined pixel by pixel, those pixels whose colors fall into the defined skin area are detected as skin tone regions. The HSV color space model is shown in Fig. 2, and the transformation from RGB space to HSV space. However, how do we define the skin area in the color space that is robust to skin colors, racial differences, and lighting conditions?

Before we discuss the distribution of skin tone color, we must know that the appearance of the skin is affected by the degree of pigmentation (varies amongst individuals and different races), the concentration of blood, and the incident light source. The combination of all these factors gives rise to a variation in skin color which spans over the range of red, yellow, and brownish-black. This corresponds to a restricted range of hue values as will be shown below. In order to gather statistics, three hundred images for each racial class of Caucasian, Asian, and African-American were used as training sets in [42]. The hue distribution for each category is given in table 2.

Caucasian		Afican-American		Asian	
m	±	*m*	±	*m*	±
25.3	6.8	8.6	8.2	28.9	5.1

Table 2. Statistics of the hue distribution categorized by race in degrees

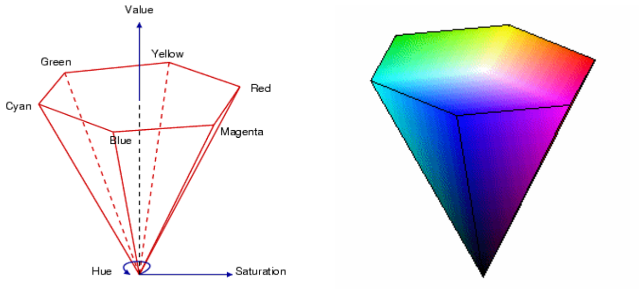


Figure 2. HSV color space model

The hue values is the main feature we use to identify skin region, after realizing the distribution of hue, we should add constrains on saturation (S) and value (V). Firstly, saturation values range from 20% to 60%, indicating that skin colors for all races are somewhat saturated but not deeply saturated. Then we notice that in the color space model of HSV in Fig. 2, as value (V) decreases, the hexagon plane also shrinks. Thus, for small V, it is meaningless to segment the hue plane, so we will define skin tone area at a high level of brightness. The actual constraints for the skin tone region and the procedure for skin detection are as follows.

$$\begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \rightarrow \begin{bmatrix} H(x,y) \\ S(x,y) \\ V(x,y) \end{bmatrix} \qquad (1)$$

The color input video is first transformed into HSV space frame by frame for skin detection. The skin region constraints are

$$S \quad 20^{0} \quad H \quad 42^{0} \quad 0.1 \quad S \quad 0.6 \quad V \quad 0.3 \qquad (2)$$

After the definition of skin tone space, pixels in the image are examined to form the skin tone map.

$$Skin(x,y) = \begin{cases} 0 & if\ HSV(x,y) \notin S \\ 1 & if\ HSV(x,y) \in S \end{cases} \qquad (3)$$

The skin map is then passed to the edge preserving binarization where additional steps will be performed to enhance the specified regions.

In Fig. 3, we will show some of the frames extracted from video clips of QCIF format and the detected skin tone regions. They are the first frame from video sequences of *carphone*, *Claire*, *foreman*, *grandma*, *miss_America*, *mother_daughter*, *Suzie*, and *Trevor* from top to bottom, from left to right. From the experiments on various types of videos, we can observe that our definition of skin tone color successfully detects most face and hand regions under indoor

Original	Skin	Original	Skin

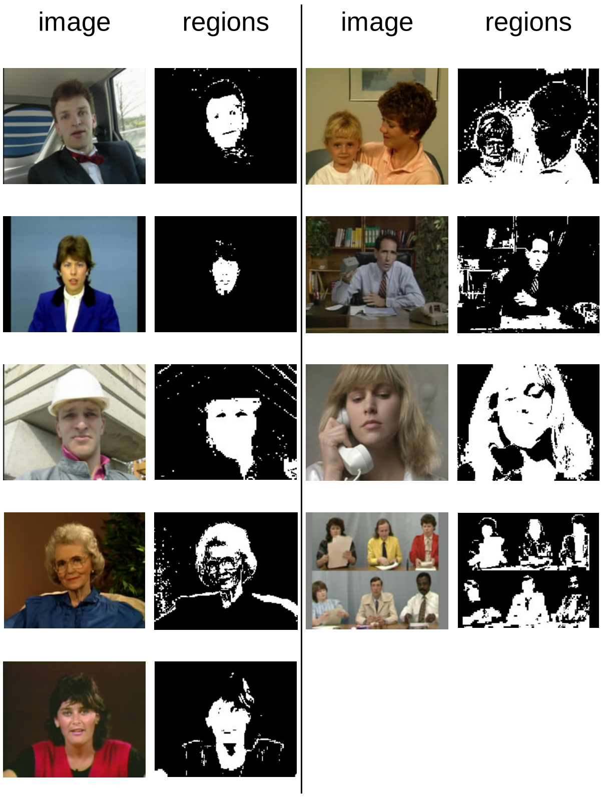


Figure 2. Skin regions detection experiments

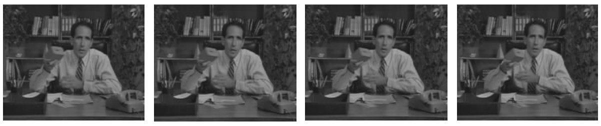
and outdoor illumination conditions. Except for video *mother_daughter*, in which the yellow incident light source has colored the whole image, most skin tone regions are detected. However, some non-skin regions are also falsely detected, such as *Suzie*'s hairs, papers and cloths in *Trevor*, and miscellaneous objects in *salesman* sequence. We will see later in the edge preserving binarization block that these false alarms have limited influences on the output images after enhancement.

Difference Update

Bi-level video can save a considerable amount of bandwidth, however, it still have temporal redundancy which represents the repeated appearance of the same content for several frames. For example, static background in the news broadcasting remains almost unchanged throughout a section of news film. Thus, if for each frame, we update only the different parts of the image compared with the previous frame, bit rate of the bi-level video can be further reduced. First, we have to identify the moving parts in two consecutive frames. A difference edge map detection method is employed [43].

$$\begin{aligned} DE_n &= \Phi\left(\left|I_{n-1} - I_n\right|\right) \qquad (4) \\ &= \overline{\ }\left(\nabla G * \left|I_{n-1} - I_n\right|\right) \end{aligned}$$

Where is the difference edge of frame *n*, it represents the moving part of frame *n* compared with frame *n-1*. Operator *G* is a Gaussian smoothing filter which suppresses the noises in the luminance difference. Here we have chosen a Gaussian filter with standard deviation of 2 for QCIF sequences. Without *G*, the difference edge will be so sensitive that tiny unperceivable illumination fluctuation will be detected. is a gradient operator, and is the non-maximum suppression to the gradient magnitude that can thin the edge. Laplacian kernel is used as the gradient operator, while is omitted in our experiments. Some QCIF sequences are tested for motion detection as follows.



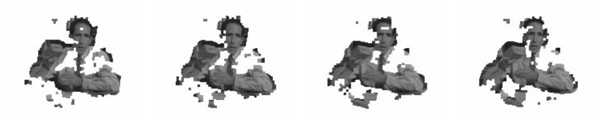


Figure 3. Moving object detection (QCIF *salesman* frame 10-13)

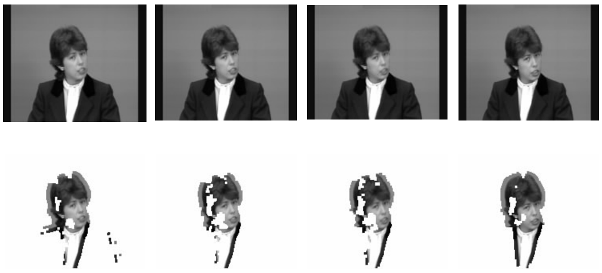


Figure 4. Moving object detection (QCIF *Claire* frame 96-99)

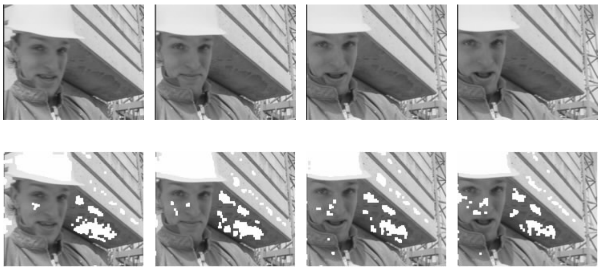


Figure 5. Moving object detection (QCIF *foreman* frame 274-277)

Fig. 3 and 4 are examples of static background videos, while in Fig. 5, a camera 'panning' action has caused both background and the foreman to move in the sense of absolute position of each frame. As the experiments show, the method of moving object detection is more suitable for sequences with static background like *salesman* and *Claire*, motion detection detects almost the full range of each frame with moving backgroud in *foreman* sequence. The motion detection for moving background videos is also proposed in [43], but it requires prior knowledge of interested objects and a lot more computation due to motion estimation. Considering the bi-level video system should have low computation demands, the motion detection for moving background is not incorporated into the system. Besides, even for videos of moving background, transmission of the sequence frame by frame still has low bit rates.

After recognizing the moving regions of each frame, we can reduce the computation requirement and bit rates by binarizing and transmit only the moving regions. However, if the exact shapes of the irregular moving regions are to be described, a lot of bandwidth will be wasted, and the system becomes more complex, too. Therefore, we define the difference update region as the smallest rectangular that contains all detected moving parts. It will take at most bits of header to describe the position and size of the rectangular, the information is also passed to the edge preserving binarization block so that only the update rectangular is binarized to save some computation amount. This procedure is illustrated as follows.

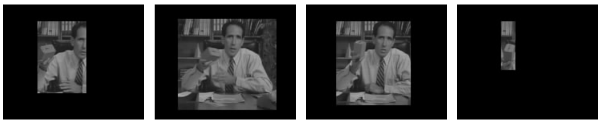


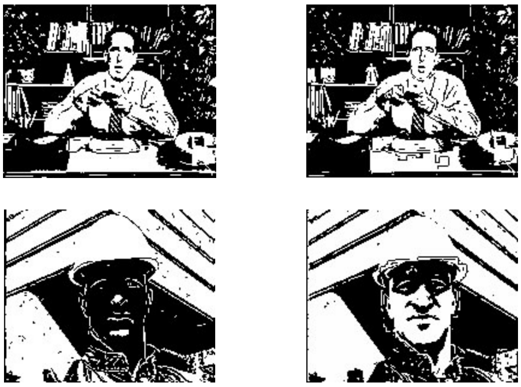
Figure 6. Update rectangular definition

(QCIF *salesman* frame 2,12,67,79)

The update rectangular can efficiently save bit rates for static backgroud videos such as *salesman* and *Claire*. But recalling from Fig. 5, the update regions for moving background frames contain almost the whole frame. Since the whole frame is transmitted with no saving in bandwidth after all, the motion detection operation becomes a waste of computation. Thus, a mechanism able to distinguish between moving and static background may be inserted before.

Edge Preserving Binarization

This function block is the edge preserving binarization with Laplacian edge detection in section 3.3, however, a few adjustments must be made to fit in the bi-level video system. First of all, the input image for this block is given by the update rectangular defined in the difference update block. Except for the first frame that has no comparison subject, this frame is binarized with full size of QCIF standard. Secondly, skin tone region information from skin detection block should be utilized to enhance the skin pixels during the binarization process. In order to achieve the skin enhancement, we first try an intuitive approach to increase the gray level of skin regions. However, this will result in additional intensity edges and affect the binarization process as shown below.



Without skin enhancement

Gray level enhancement

Figure 7. Gray level skin enhancement

From Fig. 7 we can see that although facial parts are enhanced after gray level lifting, obvious boundaries also appear around the detected skin regions. For *foreman* sequence, contour lines surround his face, as for *salesman* sequence, some false alarms caused in the skin detection block result in unnecessary noises on the table and phone. Because we lift the gray level within skin regions abruptly, the resulting edges are detected in edge preserving block and are locally binarized. If the lifting of gray level can be done gradually, this effect may disappear, but it also costs extra computation to create smooth boundaries for enhanced skin regions. Alternative approach concerns the manipulation of the threshold surface, therefore, skin tone regions enhancement is combined into the binarization block. The adaptive threshold plane for input update rectangular is obtained from eq. and is repeated below.

$$T_{adaptive}(x,y)=\begin{cases}T_{global} & \text{if } Edge(x,y)=0\\ T_{local}(x,y) & \text{if } Edge(x,y)=1\end{cases} \tag{5}$$

With the skin region *Skin*(*x*,*y*) input from skin detection block, the thresholds for skin region pixels are then lowered for enhancement.

$$T_{enhance}(x,y)=\begin{cases}T_{adaptive}(x,y) & if\ Skin(x,y)=0\\ k\times T_{adaptive}(x,y) & if\ Skin(x,y)=1\ and\ f(x,y)<T_{adaptive}(x,y)\end{cases}\tag{6}$$

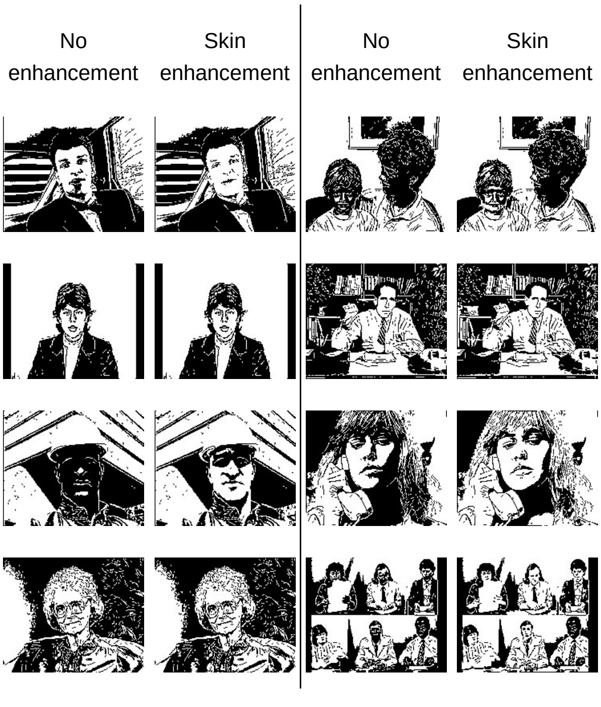
Where k value is smaller than one and is chosen as 0.7 in the system. Enhancement is executed at the originally dark skin tone pixels after adaptive thresholding, where it is necessary. The reduction of threshold values is equivalent to the increase of gray level in the binarization outputs, except that processing of threshold plane will not cause extra edge pixels and contours of skin regions at the output end. The enhanced frames are shown in Fig. 8 and 9. The thresholding for update rectangular is finally performed.

$$f'(x,y)=\begin{cases}0 & if\ f(x,y)<T_{enhance}(x,y)\\ 1 & if\ f(x,y)\geq T_{enhance}(x,y)\end{cases}\tag{7}$$

Where $f(x,y)$ and $f'(x,y)$ are the gray level and binarized images of update rectangular, respectively.

For experiments in Fig. 4.8, binarized videos originally having dark skin tone regions are *carphone*, *foreman*, *mother_daughter*, *Suzie*, and *Trevor*. In the case of *mother_daughter*, skin detection failed to include facial parts due to colored illumination as shown in Fig. 4.3. Therefore the enhancement of wrong skin tone regions is useless. To solve this problem, image processing techniques involving color invariance may be employed [44], but complexity and computation of the system will also increase. Another solution to this problem is the use of retinex filter [45] to enhance the whole image indifferently, that is, without skin tone detection. This is tried and shown below. However, the indifferent property of this method will affect perception of original images and is not adopted in the system.

In the other cases where skin detection success in containing facial regions, the shadow on the faces are lighted up, the most obvious cases are *foreman* and *Trevor*, expressions and facial characteristics can be recognized after the enhancement. As for videos that have well-illuminated skin tone regions such as *Claire*, *salesman*, *grandma*, and *miss_America*, the enhancement process keeps them intact and preserves the satisfactory results.



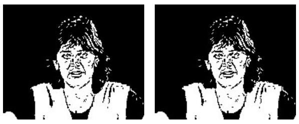


Figure 8. Threshold plane skin enhancement ~ QCIF

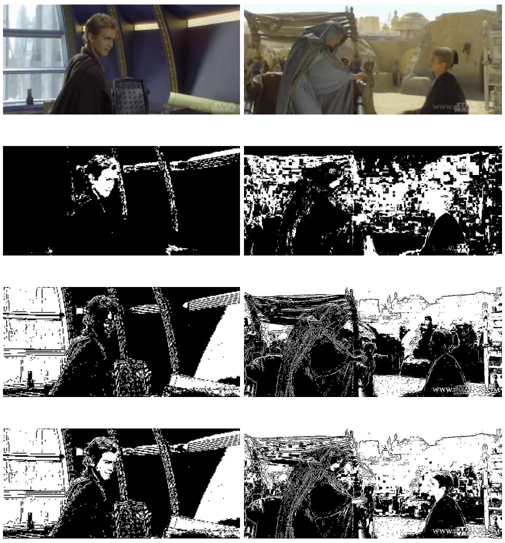


Figure 9. Threshold plane skin enhancement ~ *star wars episode 2*

Video of more complex scenes from *star wars episode 2* is tested in Fig. 9, the frames are composed of natural and computer generated images. The original frame, detected skin tine regions, binarization without and with enhancement are shown from top to bottom. Expressions on the faces are clarified for both images. Although false detection appears on the left side frame, it does not influence the output much. On the other hand, desert scene of right side image causes a lot of false alarms in skin detection, and the output is also affected. In this case, more sophisticated face allocation techniques involving shape matching may be utilized. For both images, enhancement of skin tone colors avoids the loss of essential information for human perception.

Conclusion

Application of edge preserving binarization is extended to video sequences in this report, the low bit rate nature of binary video makes it possible to transmit video sequences on the current wireless communication system. We have focused on the binarization of videos to obtain low bit rate binary videos that preserve detail information at the same time. After binarization, the raw bi-level videos can be further compressed and encoded to acquire lower bitrates. Besides the binarization process, we have introduced two extra functions, skin enhancement and difference update. They behave well in expressions enhancement and bit rate reduction for most cases. However, for some situations, such as moving background or colored illuminated sequences, more sophisticated algorithms must be employed for solutions. But under the principle of low system complexity, we do not include these algorithms. Therefore, this bi-level system works most effectively and efficiently for static back ground, well-illuminated sequences.