Independent Component Analysis and Raman Microspectroscopy on Paraffinised Non Dewaxed Cutaneous Biopsies: A promising Methodology for Melanoma Early Diagnosis

Cyril Gobinet[1], Ali Tfayli[2], Olivier Piot[2], Valeriu Vrabie[1] and Régis Huez[1]

[1] CReSTIC, URCA, Campus du Moulin de la Housse,
B.P. 1039, 51687 Reims Cedex 2, France

Unité MéDIAN, CNRS UMR 6142, URCA, 51 rue Cognacq Jay
51096 Reims Cedex, France

Abstract. This paper deals with a promising methodology for melanoma early diagnosis. Raman spectroscopy is used to record vibrational information of paraffinised tumoral tissues. Independent Component Analysis (ICA) is performed on Raman spectra to numerically deparaffinise spectra. Resulting deparaffinised spectra are used to extract discriminant information specific to malignant and benign tumors. These spectral specificities can be employed as molecular descriptors of the type of pathology. A comparison with Principal Component Analysis (PCA) shows that ICA is more suited to process this kind of problem.

1 Introduction

Cutaneous melanoma is the most severe form of skin cancers and accounts for three-quarters of skin cancer deaths. Clinical diagnosis of malignant melanoma is difficult due to its similarity to atypical benign nevi. Therefore new and efficient non-invasive tools for early diagnosis of melanomas present a crucial interest in clinical practice.

Since few years, several studies have reported the potential of vibrational spectroscopies to characterise and to differentiate cancerous from normal tissues. Raman imaging has been often used due to the fact that Raman spectra provide useful information about molecular composition of biological structures.

The paraffin embedding process enables to conserve biopsies for several years. However the use of paraffinised tissues for spectroscopical investigations remains very restricted. This is due to energetic Raman peaks of paraffin that mask important vibrational bands of the tissues in recorded spectra. Few works are related to the analysis of paraffinised tissues, but they led on chemically dewaxed and rehydrated tissues [1–2], a procedure that may induce alterations in the tissue structure, and which is time and chemical reagents consuming.

Recently Tfayli *et al.* [3] have successfully used the FTIR to discriminate between nevi and melanomas on paraffinised non dewaxed skin sections. The discrimination was based on narrow vibrational bands where the paraffin has no contribution.

Gobinet C., Tfayli A., Piot O., Vrabie V. and Huez R. (2005).
Independent Component Analysis and Raman Microspectroscopy on Paraffinised Non Dewaxed Cutaneous Biopsies: A promising Methodology for Melanoma Early Diagnosis.
In *Proceedings of the 1st International Workshop on Biosignal Processing and Classification*, pages 19-26
DOI: 10.5220/0001195500190026

In this work we propose first to numerically deparaffinise the Raman spectra by using the Independent Component Analysis (ICA). Second, we show the possibility to extract discriminant sources specific to malignant and benign tumours, sources that can be employed as molecular descriptors of the type of pathology. We also show that ICA is a more efficient technique than the commonly used Principal Component Analysis (PCA) for this kind of treatment.

2 Data acquisition and properties

Tissue sections of 10µm thick were cut from paraffin embedded biopsies (Dermatology department of Reims university Hospital). Sections were fixed on CaF_2 slides suitable for Raman analysis. Spectral images were collected by a Labram spectrometer (Dilor-Jobin Yvon, Lille, France) in a point by point mode with a 10 µm step. The light source was a titanium-sapphire laser exciting at 785 nm. In each point, the spectrum was recorded at 1305 wavenumbers covering a spectral region from 200 to 1800 cm^{-1} with a resolution of 1.22 cm^{-1}.

Due to the fact that most nevi and melanomas affect the skin epidermis in their first step of development, the analysis of each tissue is based on the processing of datasets composed of Raman spectra from the skin epidermis. The malignant melanoma and benign nevus datasets contain respectively 152 and 119 spectra. Few recorded Raman spectra are shown on figure 1.

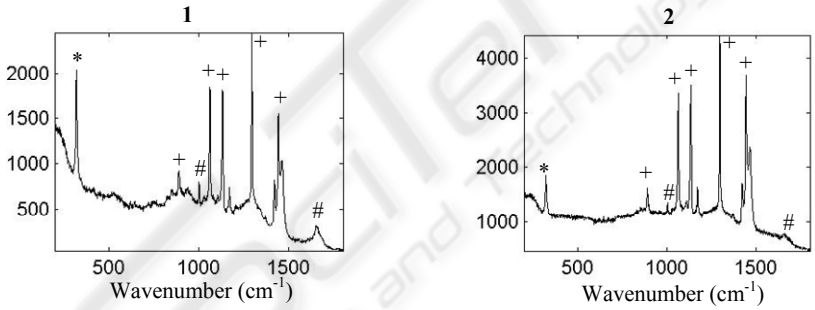

Fig. 1. Examples of recorded Raman spectra from (*1*) melanoma and (*2*) nevus. Peaks labeled by (*) are associated to the CaF_2 medium and by (+) to paraffin. The visible part of the keratin spectrum is labeled by (#).

The analysis of these Raman spectra suggests three remarks:

- whatever is the kind of analyzed tissue, paraffin and CaF_2 spectra exhibit thin energetic peaks, which are the predominant features of the recorded spectra; the contribution of keratin and melanin, which are known to be Raman active species, is not visible in the recorded spectra;
- recorded spectra are polluted by a so-called background or baseline that originates from the skin fluorescence;
- due to the spectral resolution of the spectrometer, the thin energetic Raman peaks of paraffin and CaF_2 are not aligned on the same wavenumber from a recorded spectrum to another; source separation techniques as PCA or ICA will fail to

estimate spectra of pure chemical components by computing several neighboring peaks dispatched in different sources.

Without further processing of data, no information of skin compounds can be extracted. We are thus investigated methods to extract information related to these species by examination of statistical and physical characteristics of the dataset.

The first and obvious feature is the instantaneousness data recording because the scattered light is collected by CCD detectors. Physical laws governing Raman spectroscopy mechanisms are well known to be linear. Recorded spectra thus result from a weighted sum of spectra of pure species present in the analyzed tissues. This instantaneous and linear model is:

$$X = AS = \sum_{j=1}^{M} \underline{a}_j \underline{s}_j \quad (1)$$

where X is the data matrix, S is the pure species spectra or sources matrix, and A is the mixing matrix. Each element a_{ij} of A represents the concentration of the j^{th} pure species $\underline{s}_j$ (which is the j^{th} line of S) into the i^{th} recorded spectrum. The model can furthermore be written as a sum, as shows the right member of equation (1). The j^{th} column of A, noted $\underline{a}_j$, represents the concentration profile of the j^{th} pure species. M denotes the number of sources of the model.

Spectra of the CaF_2 medium, paraffin and melanin are well known to be sparse and to possess few peaks localized in narrow bands. Peaks of one of these three species are not overlapped with peaks of the other species. Mutual independence between these pure spectra is thus a verified assumption. The keratin spectrum is not sparse and has not narrow peaks, but its smooth shape makes it to be independent of the other compounds spectra.

All conditions are combined to apply source separation techniques to the datasets.

3 Methods

To overcome previously cited problems, we propose to process data by a three step procedure.

The first step consists in suppressing the background. For each recorded spectrum, it is estimated by a five order polynomial. An asymmetric truncated quadratic cost function studied by Mazet in [4] is used to estimate the polynomial coefficients. The processed spectra are obtained by subtracting the corresponding baseline from each recorded spectrum. An example is given on figure 2.

The alignment of CaF_2 and paraffin peaks is realized in the second step. This part consists in upsampling spectra in spectral bands where a peak is localized, in computing the shift between a reference spectra peak (commonly the first spectrum recorded on each tissue) and the other spectra peaks, in shifting back the peaks in order to align their maximums, and finally in downsampling spectra. This alignment is commonly encountered in geophysical signal processing [5].

The last step corresponds to the elimination of CaF_2 and paraffin influence. Two different approaches are possible, the Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The use of PCA is motivated by its common application to biological and biophysical datasets [2]. PCA is searching for

statistically decorrelated sources (pure species spectra) that respect the linear model described above. The decorrelation is only a second order independence, so estimated spectra may be a linear combination of pure species spectra. In ICA methods [6, 7], the decorrelation assumption is replaced by the statistical independence of unknown sources (Raman pure species spectra in our case). This hypothesis is in respect with the pure species spectra characteristics mentioned in section 2. The JADE algorithm [7] was used here to estimate the components model. ICA has proved its efficacy to a wide class of applications [8]. Note that, as usual, application of PCA or ICA is preceded by centering of data.

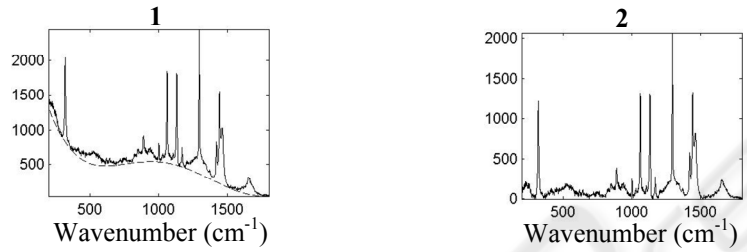

Fig. 2. Background removal from a recorded spectrum: (1) estimation of the baseline by a polynomial of order 5 (*dashed line*) from a recorded spectrum (*solid line*); (2) removal of the baseline by subtraction

4 Results and discussion

4.1 PCA

After the alignment and centering steps, PCA is applied to datasets. The two first principal components are depicted on figure 3.

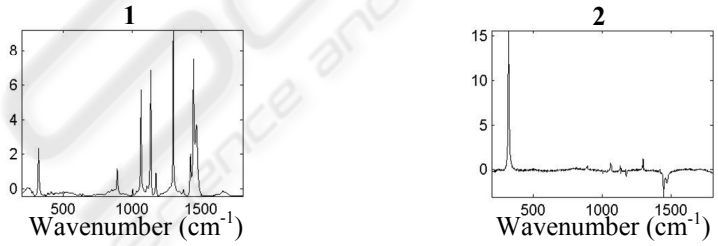

Fig. 3. The two first principal components estimated on benign nevus

Even if the study of the principal components may lead to discrimination between tissues, pure species spectra are not well identified. Paraffin and CaF_2 spectra cannot be exactly subtracting. It is shown that the first estimated spectrum at the left of the figure has its peaks well oriented. The second one at the right of the figure exhibits peaks oriented to opposite directions. This is physically unrealistic. Moreover, concentration profiles of these principal components exhibit negative and positive

values. To feat to reality, only totally positive concentrations are admitted. Moreover, the spectra estimated by PCA are still linear combinations of pure species spectra as can be observed in figure 3 where influence of CaF_2 and paraffin are mixed.

4.2 ICA – 3 components

This incorrect estimation motivates the use of ICA. The JADE algorithm [7] was used to estimate for each kind of tissue a three components model predicted by the PCA analysis. Estimated pure species spectra, corresponding to sources, are depicted on figure 4 for a nevus.

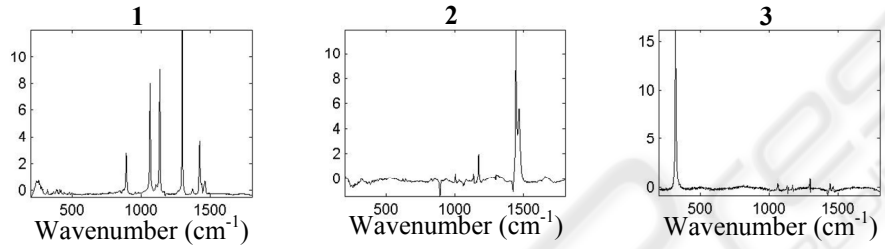

Fig. 4. Independent components estimated by a three sources model on the benign nevus

Identical results are obtained for a melanoma. The first source is associated to a part of the paraffin spectrum. The second source corresponds to another spectral band characteristic of the paraffin spectrum. The third source is the spectrum of the CaF_2 medium. A first conclusion is that paraffin acts differently with the underlying tissue in function of the considered spectral band. A second conclusion is that paraffin and CaF_2 spectra are too much energetic compared to keratin or melanin spectra. These two remarks suggest decomposing spectra with more than three sources. The number of sources must be sufficient to decompose paraffin in independent behavioral spectral bands and to estimate the poorly energetic keratin and melanin spectra.

4.3 ICA – 5 components

A four components ICA model lets keratin spectrum mixed with a paraffin source. A five components model leads to a well decomposition of paraffin (three sources) and keratin (one source). Nevus estimated independent components are shown on figure 5. Similar results are obtained for the melanoma. Only the keratin source differs from one tissue to another. Sources variance is fixed to unity. Paraffin spectrum has been decomposed in three independent spectral bands. As in the model with three sources, the CaF_2 medium conserves its unique Raman peak spectrum. The last source is similar to the known spectrum of keratin.

Remark 1. Melanin is *a priori* supposed to be a Raman active species, but its spectrum is hidden by spectra of paraffin and CaF_2. Even if the number of independent components is increasing, no matching with this spectrum was found.

Remark 2. The accuracy of spectral decomposition by ICA is demonstrated thanks to the positivity of the estimated mixing matrix. Furthermore concentrations maps can

be organized to show each species repartition in tissue.

The decomposition in several sources of paraffin spectrum does not handicap the interpretation of results because paraffin is considered as a polluting component in this application. The interesting information is the keratin spectrum.

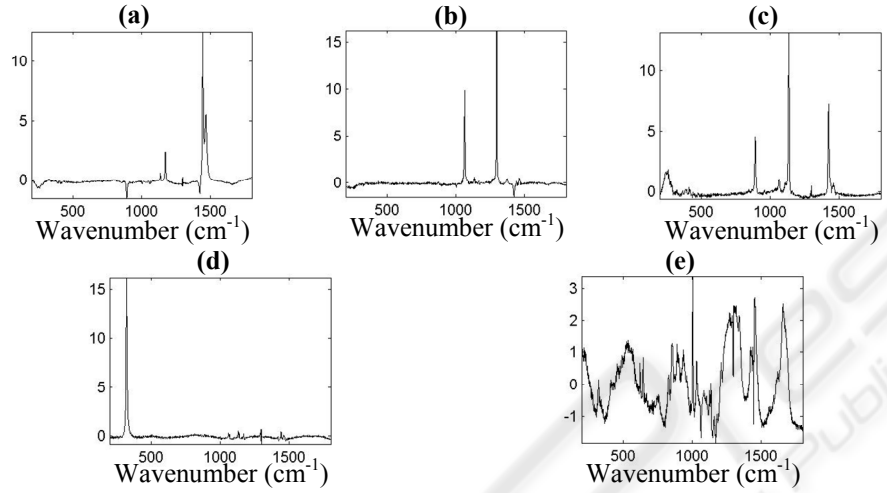

Fig. 5. Independent components estimated by a five sources model on the benign nevus. (a), (b) and (c) independent components associated to paraffin. (d) independent component associated to CaF_2. (e) independent component associated to keratin

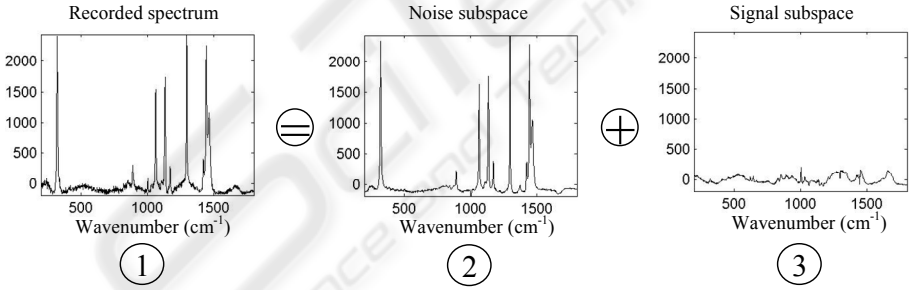

Fig. 6. Decomposition of initial data space into a noise subspace and a signal subspace. (*1*) A recorded spectra lying in the data space. (*2*) Its noise part lying in the noise subspace. (*3*) Its signal part lying in the signal subspace.

4.4 Subspace representation

To illustrate the efficacy of numerical deparaffining, let us consider that original data can be decomposed in two subspaces. The first one is called the noise subspace and is composed by uninteresting information, e.g. paraffin and CaF_2 spectra. The second one, made up by the keratin spectrum, is the signal subspace. It contains useful information to discriminate the kind of tissue. The original data matrix X can be written as the sum of the noise subspace X_b and the signal subspace X_s:

$$X = X_b + X_s \,. \tag{2}$$

Each subspace is defined by:

$$\begin{aligned} X_b &= \underline{a}_{para_1} \underline{s}_{para_1} + \underline{a}_{para_2} \underline{s}_{para_2} + \underline{a}_{para_3} \underline{s}_{para_3} + \underline{a}_{CaF_2} \underline{s}_{CaF_2} \\ X_s &= \underline{a}_{kera} \underline{s}_{kera} \,. \end{aligned} \tag{3}$$

To understand these concepts of noise and signal subspaces, let us illustrate by an example. A spectrum of the nevus is shown on the left of figure 6. Thanks to ICA, it is decomposable in two spectra. The one at the middle of the figure is lying in the noise subspace, while the second at the right in the signal subspace. Note that this last one is just a scaled version of the spectrum of keratin in figure 5(e). We can notice that the signal subspace is not very energetic compared to the noise subspace.

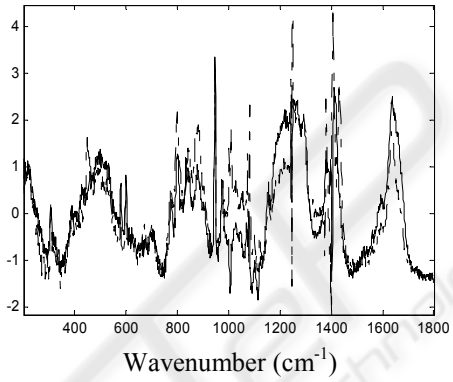

Fig. 7. Keratine spectrum estimated on a benign nevus (*solid line*) and on a malignant melanoma (*dashed-dotted line*)

4.5 Keratin spectrum discrimination

A comparison of keratin spectra estimated for a malignant melanoma and a benign nevus can be done, as shown on figure 7. The sources obtained from ICA show visible differences between nevi and melanomas. Such differences are visualised with the changing intensity ratio of the Fermi doublet bands on 850 $cm^{-1}$and 830 cm^{-1} for melanomas it is around 2.5 while it is only 1.6 for the nevi. Such changes could inform us about the state of the phenylic cycle in the tyrosine residu and the type of resulting molecular bands (intra- or inter-).

Secondary structure variations are marked by a predominance, in the melanoma source, of the α helix vibrations (1650cm^{-1}) in the amide I band. Similar information can be obtained from the high intensity of the band on 934 cm^{-1} characterising the C-C stretch in the α helix. On the other hand, the nevi source represents a shoulder band at 1670 cm^{-1} revealing a more important contribution of the β sheet conformation.

The differences in the secondary structure can be quantified by the decomposition of the amide I band by creating spectral models with Gaussian-Loretzian functions.

The same information can be obtained from the changes of the amide III band, and from the intensity of the band on 901 cm^{-1}.

5 Conclusion

When a skin sample is paraffinised, the direct analysis of recorded Raman spectra is not possible because of the predominant intensity of paraffin spectrum over the other compounds spectra. Thanks to the mutual independence of spectra of these species, ICA is applicable and estimates physically meaningful sources.

A first conclusion is that paraffin spectrum can be decomposed in three independent behavioural spectral bands. It means that the underlying tissue is more or less reacting with some spectral bands of the paraffin spectrum. A second is that melanin spectrum is not visible when paraffinised tissues are considered because of the too energetic peaks of paraffin and CaF_2. Third, more sources than suggested by PCA must be employed in order to reveal the low energetic spectrum of keratin.

The last and important conclusion is that estimated keratin spectra of paraffinised benign nevus and malignant melanoma contain a large amount of information. Little spectral differences between these spectra lead to the identification of the kind of analysed tissue. Molecular descriptors of the type of pathology have been found.

References

1. Gniadecka, M., Wulf, H., Mortensen, N., Nielsen, O. Christensen, D.: Diagnosis of basal cell carcinoma by Raman spectra. Journal of Raman Spectroscopy, Vol. 28 (1997) 125–129
2. Sigurdsson, S., Philipsen, P., Hanson, L., Larsen, J., Gniadecka, M., Wulf, H.: Detection of skin cancer by classification of Raman spectra. IEEE Transactions on Biomedical Engineering, Vol. 51 (2004) 1784–1793
3. Tfayli, A., Piot, O., Durlach, A., Manfait, M.: Discriminating nevus and melanoma on paraffin embedded skin biopsies using FTIR microspectroscopy. BBA General Subjects, accepted manuscript
4. Mazet, V., Carteret, C., Brie, D., Idier, J., Humbert, B.: Background removal from spectra by designing and minimising a non-quadratic cost function. Chemometrics and Intelligent Laboratory Systems, Vol. 76 (2005) 121–133
5. Vrabie, V., Le Bihan, N., Mars, J.: 3D-SVD and partial ICA for 3D array sensors. In: 72nd International Conference of the Society of Exploration Geophysicists (SEG'2002). Salk Lake City, USA.
6. Comon, P.: Independent component analysis: a new concept?. Signal Processing, Vol. 36 (1994) 287–314
7. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. IEE Proceedings F, Vol. 140 (1993) 362–370
8. De Lathauwer, L., De Moor, B., Vanderwalle, J.: Fetal electrocardiogram extraction by blind source subspace separation. IEEE Transactions on Biomedical Engineering, Vol. 47 (2000) 567–572