SOFTWARE APPLICATIONS FOR OVERSAMPLING OF TRANSIT CANDIDATES

Céline G. Quentin[1], **Pierre Barge**[1], **Raphael Cautain**[1], **Jean-Charles Meunier**[1], **Claire Moutou**[1], **Renaud Savalle**[1], **and Christian Surace**[1]

Laboratoire d'Astrophysique de Marseille, Traverse du Siphon, B.P.8, F-13376 Marseille Cedex 12, France

Abstract

In the Exo-field, the standard sampling of 12000 light-curves is 8.5 minutes but a finite number of them (1000 at maximum for a given field of view) can be oversampled to 32 seconds. This possibility was planned to enhance the scientific impact of the mission with the aim to get additional information on the shape and timing of the planetary transits. The targets to be oversampled are selected from a ground-based analysis of uncompletely corrected (preliminary) N1 data and sorted in a list that can be changed during the data acquisition. At the beginning of a run the 1000 oversampling possibilities are occupied either by stars known to host a planet (from ground based observations) or by reference stars appropriately chosen within the Herstzsprung-Russel (HR) diagram. We present the software used for early detection in raw light-curves and the way results are stored and sorted to build up the oversampling list that is loaded on board the satellite. This list can be changed once a week during the operations while new sets of data are acquired. The whole software and data management developed to decide which target stars merit or not oversampling will be called "Exowarning pipeline".

1. General framework

Observations with CoRoT will produce some 60000 light-curves during the long runs (150 days) and some 60000 other light-curves during the short runs (20 days). Over 150 days the expected planetary signals have periods less than 3 months and durations that ranges from 1 hour to 0.5 days. The standard sampling of these light-curves is 512 seconds when the targets are monitored in the regular mode. The sampling can be reduced to 32 seconds for a finite number of targets, at most 1000 in a given field of view. The list of oversampled targets is modified on board, and is triggered by telecommands sent from the ground. The decision to change or not the sampling mode is taken on ground after a selection procedure using specific softwares and list management tools. Targets which require oversampling are chosen among those in which the raw light-curves are found to contain transit-like events. Specific algorithms are necessary to work with high noise level and to look for theses events in the short data sets delivered during the operations. The decision procedure for oversampling results from the management of lists of targets sorted following an estimated confidence level. In this paper are presented: (i) the various algorithms developed for the processing of the data and the production of lists of candidates (targets whose light-curves contain transit features), (ii) the criteria and the selection procedures used to produce oversampling lists, (iii) the management during the operations of the data, the intermediate results and the list of candidates.

2. The operational constraints

The goal of the exowarning pipeline is to deliver to the CoRoT Mission Center (CMC), once a week, a list of targets whose light-curves have the most relevant indication of planetary transits. The decision procedure inevitably requires the use of transit detection algorithms. However, the detection will be performed on preliminary data (N1 data) issued from the correction pipeline after only a first level of instrumental corrections. Completely corrected data (N2 data) will be produced at the end of a run, after updated processing and scientific validation. The detailed analysis of the light-curves for detecting planetary transits and producing scientific results will be made on the less noisy N2 data. At the end of a run some information on the intrinsic variability of the targets will be also available and will be very useful to improve detection. As a consequence, the detection algorithms of the exowarning pipeline will be less efficient than algorithms working on the N2 data. In fact, a number of preprocessing will be necessary before detection algorithms can be successfully applied to the N1 data.

Unlike standard detection, early detection is made the operations and must satisfy two requirements: (i) to work on light-curves of short duration at the beginning of a run; (ii) to search for transit periods that can be long (up to 50 days) as compared to the total duration of the light-curves.

For this reason we propose an alternative method that can detect individual events in a light-curve. In the limit of the large transit numbers this method can be compared to classical methods such as the Box-fitting Least Square (BLS). Of course, transit detection algorithms may produce false alarms and it is necessary to estimate a confidence level for each detection. The confidence levels will be very useful to sort the detected events.

Proc. 1st *Eddington* workshop *"Stellar Structure and Habitable Planet Finding"*, Cordoba, 11–15 June 2001 (ESA SP-485,

Detections can be also ambiguous due to possible confusion between planetary transits and eclipsing binaries. In a number of cases such ambiguities can be removed when the secondary transit is very different from the primary transit. So, the most obvious ambiguities should be removed with complementary analysis of successive transit features in a light-curve.

Finally, due to the amount of data to be processed in a short time (basically less than a week) we also need fast and optimized production softwares. In addition the pipeline must be flexible enough to easily accommodate changes in the procedures and information management during CoRoT's operations.

3. Management and transfer of the data

The exowarning pipeline will run using N1 data received at "Laboratoire d'Astrophysique de Marseille" (LAM) on a weekly basis. Due to the short delay available for the processing of the data, the management of intermediate results and the transfer of the oversampling list, it is crucial that N1 data be stored in a local database. Therefore, we developed an operational Relational Database Management System (RDBMS). Formats and design of the database are based on N1 data specifications. Data exchanges, based on File Transfer Protocol (FTP) are implemented using the "Systeme d'Echange de Fichiers" (SEF) operated by CNES. Input data consists of several types of FITS files containing binary tables extensions. Files and oversampled data are identified using files naming rules. They contain chromatic and monochromatic light curves but also background information, offset images and thumbnails, calibration such as sky background and jitter correction surfaces. The data will be inserted in the database in large (file based) transactions and validated using the strong typing and reference integrity rules of the RDBMS.

On the other hand, in output, the oversampling list will be put in a XML file that should be sent automatically to the CMC using the SEF. This file is identified by its generation date and contains sorted element composed by the CoRoT-ID of the stars to be oversampled, and a tag tracking the source of the choice: Core program (possible transit candidate) or Additional program or Initial sequence based on statistic of the star targets.

4. The production database

A reliable and scalable central repository is developed at LAM for the exowarning pipeline. This database, implemented with the Oracle software, will be updated weekly at each release of a new N1 data set. It will contain all the products necessary for data preprocessing, transit detection and list management; it will also contain intermediate results and by-products. The database should be scalable to accommodate at least 3 years of continuous observations of more than 120,000 stars representing about 1Gb of data per week. A grid-based replication architecture will be used to ensure the reliability of the data. Since all the N1 data will change every week, the production database will implement time travel to allow the pipeline to be rerun using previous datasets.

The pipeline can access the data via business objects written in IDL (Meta Language for visualization and analysis data) and Java. They hide the complexity of the database and provide a high level data interface to the pipeline. Instead of coding complex SQL queries, developers can start software integration early and independently of database development.

Early in the design, we stressed the importance to link the production database with the EXODAT database (more details on Exodat section - Deleuil in this book), a complete dataset of astrophysical parameters for stars in the exofields collected for the preparation of CoRoT observations. This link will permit to get the most complete information on the target stars but also to remove a number of ambiguities due to background eclipsing binaries.

The production architecture includes: a database server with disk redundancy, an application server running the pipeline, a file server for raw data archival and database backups. All the machines are running on the same 10GB ethernet network (see Figure 1).

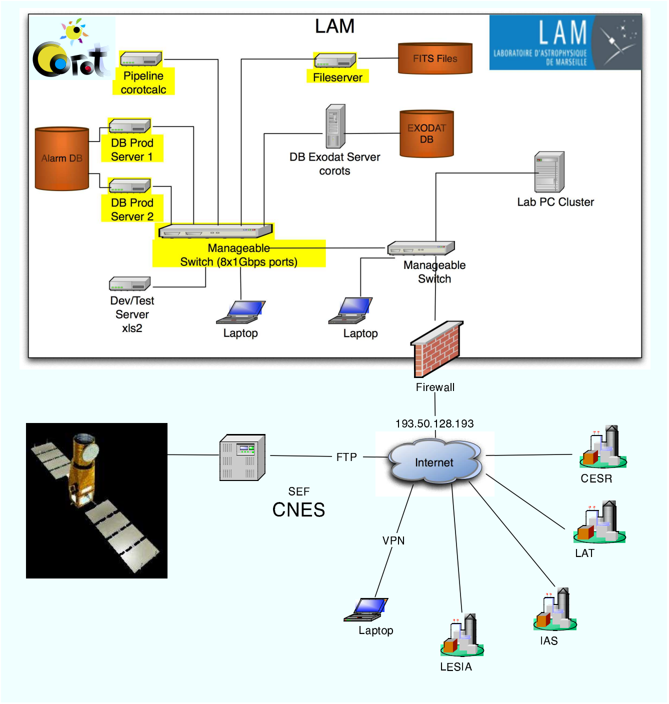

Figure 1. The core architecture is based on a 10Gb ethernet network connecting a fileserver, a computing server and the replicated database system within a VLAN. Data retrieval from the CNES SEF site is achieved over the internet.

5. Processing of the data

The N1 data in which planetary transits may be present will be corrected for a number of instrumental noises (offsets due to the electronics and thermic noises), background contributions (zodiacal light and earth scattered light from the Earth). However, the N1 light-curves will still contain a number of residual peaks due to the South Atlantic Anomaly (SAA). Other uncertainties in the corrections of scattered light and others orbital noises will also need to be analyzed in greater detail. So, before starting the detection procedures we will apply on the light-curves a number of detrending algorithms.

5.1. The detrending tools

A moving box filtering will be used to remove the of the SAA peaks. It consists in averaging over a few orbits the signal contained in a box and to remove values greater than a given threshold equal to a few percents of the mean value. To follow the slow variation of the signal, we can use a time-frequencies analysis. First we apply a Fast Fourier Transform to the light curves and, working in the frequency domain, filter the contributions outside the transit range. At last, we apply the gauging filter described in Guis and Barge, 2005 in order to reduce the noise level and to improve morphological transit detection. Based on a 1D-morphological filtering (more details in Serra and Soille, 1994), this filter uses aperture and closure operators with an elliptic structuring element (or gauge); the gauge's half-width determines the time sampling of the filtering . The various detrending steps of the N1 data are illustrated in Figure 2. Theses tools will evolve during the operation as soon as we learn from the instrumental environment.

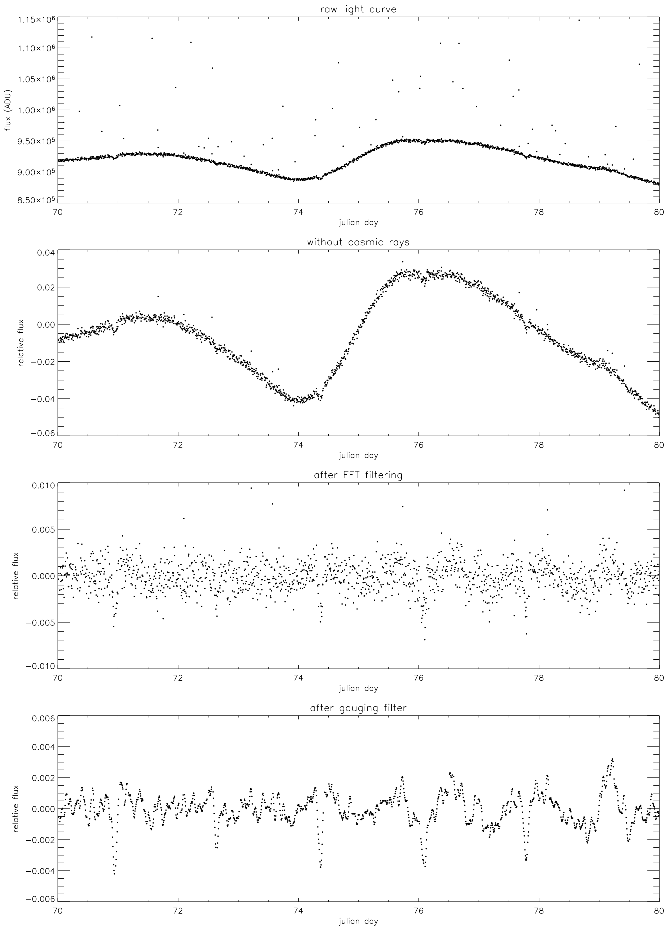

Figure 2. Example of simulated N1 data containing transit events. The light-curve before (top) and after (bottom) the denoising steps (respectively). The moving box is applied first to remove the residual peaks of the SAA, then, the slow variations of the signal are reduced thanks to a Fourier analysis, finally the gauging filter removes residual orbital noises to exhaust possible transit signals.

5.2. Transit feature identification

Our goal is to detect transit signals along the data flow, to estimate the confidence we have in the detection and to start oversampling as soon as possible for the best candidates. To this end we developed two detection algorithms which are, in fact, complementary from one another. The first one is designed to look for single events in light-curves of short duration, the second one focuses on the detection of periodic events with large transit number.

The Morphological Individual Detector (MID, presented in Quentin et al., 2005) was developed to detect single transit events and estimate their depth ($\Delta F/F$ the relative flux) and duration. The detection of transit like features in a light-curve is performed individually on a T blocks of data of short duration (see Guis and Barge, 2005 for the principle of the method). This can be done using watershed segmentation of the signal and its opposite (a morphological analysis described by Beucher, 1994). The results is a measurement $X(T) = [D, W, S]$ of three criteria: depth, width, and surface of an event at the block data's date. Then, each measurement $X(T)$ on the signal can be classified into two categories (or clusters): noise X_n or detected event X_d. Points in the opposite signal help to supervise the clustering, as it enhances the reference cluster or "noise cluster". A Ratio for Detection Efficiency (RDE) is defined as the ratio of the difference to the sum of two quantities: the mean depth of a detected event $< D_d >$ and the mean depth of the associated noise $< D_n >$.

$$RDE = \frac{< D_d > - < D_n >}{< D_d > + < D_n >} \tag{1}$$

RDE decreases with the signal to noise ratio (SNR), and increases with the transit depth. Using RDE to sort the light curves, favours the deepest transit, and reduce the level of false detections, due to a bad SNR.

The Box fitting Least Square (described by Kovács et al., 2002) is a standard algorithm based on a fit with a square-well transit model and on a folding of the signal. This algorithm is a very efficient one (see Tingley, 2003) even for low SNR, but as far as the period of the transit is shorter than two or three times the duration of the observations. We used the Signal Detection Efficiency (SDE) to sort the detection. Kovács et al., 2002 defined it as the normalized Signal Residue:

$$SDE = \frac{SR_{peak} - < SR >}{\sigma(SR)} \qquad (2)$$

where SR_{peak} is the highest value in the BLS spectrum, $< SR >$ the average of the spectrum and $\sigma(SR)$ its standard deviation.

5.3. Search for periodicities

Applied on a light-curve MID may provide a set of transit like features with their associated dates. This set of features can result from the detection of a periodic signal in which some transits have been missed or, on the contrary, have been erroneously detected. The second stage in our analysis is to look at the possible periodicities within the set of the detected features and to test the identified periods, confirming or infirming them. A specific algorithm was developed to that purpose: the Event Periodicity Finder (EPF). In a first step, EPF provides a set of candidates (periods, epochs) fitting at best the set of the estimated dates and widths. Periods are chosen through a floating divider algorithm (FDA), which divides interval durations without any folding. An example is given in Figure 3. FDA enables to find a period even if many transits in the sequence are undetected. Each candidate have to match, at least, 3 or more events. EPF is forming classes containing similar candidates, then it retains the classes that match with the largest number of events.

The second step is a standard folding of the signal in which the epoch, the period and the width of the events are approximately known. So, the folding is performed after a windowing of the signal using the Phase Dispersion Minimization introduced by Stellingwerf, 1978, as illustrated in Figure 4. This method, called hereafter WPDM, allows to reduce the noise contribution and to use a predictive model of the periodogram's peak. EPF can also estimate how the periodogram's peak is fitting to its model and provides a confidence level. This level is low when the candidates are built with one or more false detections.

5.4. Removing the most obvious ambiguities

When SNR is high enough, it becomes possible to discriminate transiting planets from eclipsing binaries. To this end we use the EPF algorithm that provides the period and epoch (P,E) of an event. The signal is folded according to the same width of the event but changing (P,E) into either (P,E+P/2), (2P,E) or (2P,E+P), as shown in Figure 5.

A confidence level is computed in two cases : (i)the presence of an unseen transit at (P, E+P/2); (ii) a significant difference in the depth of the two events (2P, E) and (2P, E+P). These confidence levels are homogeneous to a SNR in both cases. Sensitivity is not optimal, as far as WPDM is devoted to detection and not to measurement; it does not take into account any astrophysical constraints. Nevertheless, the method can be easily implemented in our detection procedure and allow to test early in the process a number of ambiguous configurations. For more precise studies, other specific tools can be developed to handle short periods binaries or to systematically filter long period tidal effects.

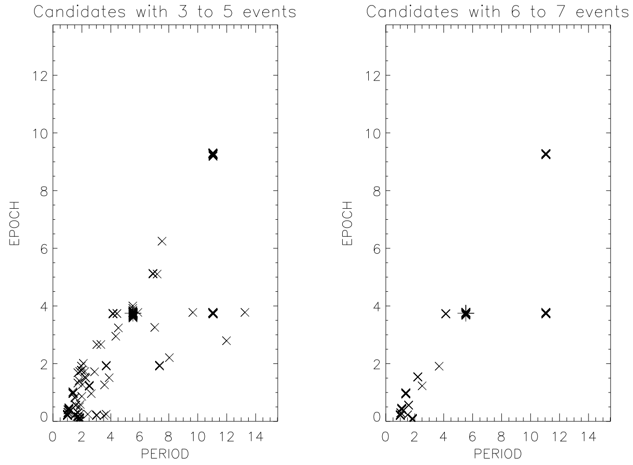

Figure 3. Possible candidates (period, epoch) matching a number of features that ranges from 3 to 7, among a total of 30. Candidates matching many features (right) are less frequent than others (left). The plus sign indicates the true events: (5.52, 3.75).

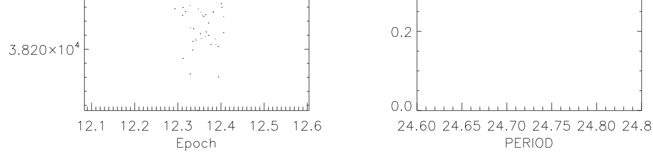

Figure 4. Example of WPDM folding. The width of the Window is 5, and the bin size is 0.25 (unit : estimated transit width). The periodogram enables to improve the determination of the period.

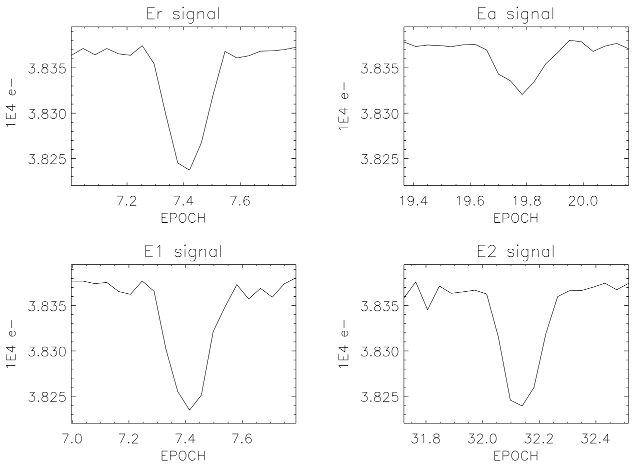

Figure 5. The folded signals at the periods and epochs Er = (P, E), Ea = (P, E+P/2), E1 = (2P, E) and E2 = (2P, E+P). A secondary transit is found with a relative depth of 0.38 and a confidence level of 4.2. Transits at E1 and E2 are found to have similar shapes: the relative difference is 0.06 but with a confidence level of 0.62.

6. Light-curves simulation

To test and estimate the robustness of the algorithms (MID, BLS, WPDM) developed to identify transit-like features and to find their associated period, we used simulated light curves. The capacity to discriminate a transit candidate from the signal of simple eclipsing binaries is also evaluated using some test cases.

Different sets of synthetic light-curves have been constructed by combining several components: (1) the output of the instrument model in order to account for realistic noises affecting CoRoT light curves), (2) a modeling of the stellar micro-variability , (3) a simulation of the possible planetary transits, (4) a simulation of the signal of an eclipsing binary or a variable star.

A first set of light curves were produced in the framework of the blind test exercise where a few tens of events were included in a sample of 1000 simulated curves. This exercise allowed a first tuning of various detection methods and a refinement of the detrending algorithms in the CoRoT context. Details on the light-curve simulation are given in Moutou et al., 2005. The temporal sampling of the final light-curves is 8 minutes, with a duration of 150 days, as for the CoRoT long observing runs. The content of a simulated light-curve is illustrated in Figure 6. The package of 1000 light-curves were supplied to five independent detection teams with no information on the way they were simulated; neither the number of hidden planets nor the nature of injected noise sources were known by the detecting teams (this data set is called BT1 in the following). This exercise also permits to determine how manage the detection and classify them by their confidence levels and priority.

Then, another sample of 850 light-curves were created in a similar way, all including planetary transits. The transit parameters were chosen to regularly sample the range to which CoRoT data will be sensitive, i.e. depth 10^{-4} to 5.10^{-2}, duration 30 to 360 min, and orbital period 3 to 50 days. Three levels of activity were included: no activity, faint and strong. The single magnitude 14 and spectral type G was used in this data set. Such collection of simulated light curves was used to learn how to tune at best the detection and detrending algorithms in the framework of the oversampling mode. It also enables to learn about the detection capability.

Finally, a third data set was settled, containing three-color light-curves for each star. It was used in the second blind test exercise performed mainly to test the characterization capacity when using color information. Chromatic dependence of instrumental noises, stellar micro-variability, transits of eclipsing binaries and star-planet systems was included. All light-curves contain a transit event (with various level of difficulty for the detection) and the objective is to identify the true origin of the event. These data (BT2 sample) was used in addition to the previous ones, in order to test the detection and the discrimination method.

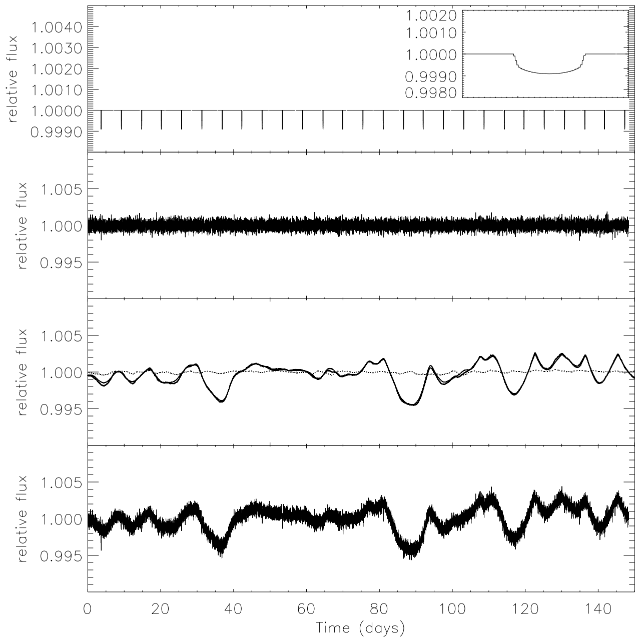

Figure 6. Example of light-curve with all contributions: from top to bottom, the transit of a planet (with a zoom around one of the transits), the instrumental noise on a 13 magnitude G-type star, the stellar micro-variability of two stars included in the CoRoT mask (G5 and G8 types respectively for the bright (thick) and the faint (thin) star, the magnitude difference being 5.2), and the final light-curve. All intensities are normalized to unity; the mean level is 2.10^5 photo-electrons per exposure.

7. Oversampling list management

During a run of observation, the exowarning pipeline will provide the Oversampling List (the list of targets that require oversampling, hereafter OL) once a week, and will also manage possible changes in the OL. However, at the beginning of a run, the delay before the first alerts (with change in the OL) could be a bit longer than a week due to the time necessary for tuning the various algorithms and waiting for the first detections. This time lag before the first alerts will be used for another purpose: oversampling of peculiar targets selected by the various scientific programs of the mission and sorted following specific priorities. To proceed this way will optimize the use of CoRoT's oversampling capability. On the exoplanet side initial OLs will contain: (i) all the stars known to host a planet from the preliminary ground based surveys; (ii) a sample of stars chosen within the HR diagram (these stars could be used to calibrate the stellar variability following the spectral type and to tune the transit detection algorithms).

Among the one thousand windows that will be free for oversampling, some of them will be entirely devoted (all along the 150 days of observation) to peculiar scientific programs. They will be free to choose their targets before the beginning of a run of observation, and will provide the list of targets to the exowarning team. For the other 950 windows the priority is attributed to the Exoplanet Core program, except if unused as may occurs at the beginning of a run where priorities may change for a while.

During the observation the OL will change according to a number of criteria chosen a priori and to the values of parameters computed along the exowarning pipeline. The criteria for the sorting of the OL are the following: (P0) highest priority, attributed by a scientific staff to a number of unambiguously defined peculiar events; (P1) candidates detected by our two methods (BLS + MID/EPF), both; (P2) periodic events detected by the standard BLS method; (P3) events with low transit number detected with MID and EPF; (P4) single transit features detected with only MID; (P5) confirmed eclipsing binaries.

8. Conclusions

The software developed to implement the oversampling capacity of the exoplanet channel requires different algorithms for: (1) detrending the raw N1 data, (2)identifying transit-like features in the detrended light-curves; (3) searching for possible periodicities between the detected features; (4) removing the most obvious ambiguities; (5) estimating a confidence level in the various detections. Two different detection procedures are used in paralell to improve the confidence level in the detected events: the Morphological Individual Detector (MID) and the standard Box-fitting Least Square (BLS). The two algorithms are complementary from one another since the first one is looking for single features in sliced light-curves whereas the second one is looking for features matching a signal, folded at a period that is folded a priori. MID is well suited for small transit numbers and BLS is a more standard algorithm, very efficient for the large transit numbers. The various algorithms have been tested successfully against a large set of simulated light-curves produced with CoRoT's instrument model and for different sources of noise. The lists of transit candidates resulting from the detection procedures are sorted as a function of confidence level and periodicity.

A crucial point is also the management and the transfer of the data during the processing steps and the production of the oversampling lists. Indeed, oversampling has to be trigerred during the operations what requires strong constraints on organization and efficiency. A database is developed to gather all the products necessary for data preprocessing, transit detection and list management; it will make easier the various aspects of the pipeline . It will also contain intermediate results and by-products.

References

Beucher, S. (1994). Watershed, hierarchical segmentation and waterfall algorithm. In Serra, J. and Soille, P., editors, *Mathematical morphology and its applications to image processing*, pages 69–76. Kluwer Academic Publishers.

Guis, V. and Barge, P. (2005). An Image-Processing Method to Detect Planetary Transits: The "Gauging" Filter. *PASP*, 117:160–172.

Kovács, G., Zucker, S., and Mazeh, T. (2002). A box-fitting algorithm in the search for periodic transits. *A&A*, 391:369–377.

Moutou, C., Pont, F., Barge, P., Aigrain, S., Auvergne, M., Blouin, D., Cautain, R., Erikson, A. R., Guis, V., Guterman, P., Irwin, M., Lanza, A. F., Queloz, D., Rauer, H., Voss, H., and Zucker, S. (2005). Comparative blind test of five planetary transit detection algorithms on realistic synthetic light curves. *A&A*, 437:355–368.

Quentin, C. G., Cautain, R., and Barge, P. (2005). Improving transit detection with collective light curves information. In Gabriel, C., Arviset, C., Ponz, D., and Solano, E., editors, *ASP Conf. Ser. 285: Astronomical Data Analysis Software and Systems XV*.

Serra, J. and Soille, P., editors (1994). *Mathematical morphology and its applications to image processing*. Computational Imaging and Vision. Kluwer Academic Publishers, Dordrecht.

Stellingwerf, R. F. (1978). Period determination using phase dispersion minimization. *ApJ*, 224:953–960.

Tingley, B. (2003). Improvements to existing transit detection algorithms and their comparison. *A&A*, 408:L5–L7.