Oscylacje wapniowe w symbiozie z roślinami. Program rozpoznawania typu sygnału.

Andrzej JARYNOWSKI*

Słowa kluczowe:
oscylacje wapniowe, metody bayerowskie, rozpoznawanie wzorców.

Zrozumienie symbiozy roślin strączkowych z grzybami (mikoryza), czy bakteriami (ryzobia Nod) jest jednym z najważniejszych biotechnologicznych wyzwań w polepszeniu efektywności rolnictwa (Hazledine 2008). Naszym obszarem dociekań była zmiana poziomu stężenia jonów wapnia w wyniku interakcji symbiotycznych (mierzona póki co tylko w dwóch ośrodkach na świecie, a nasza praca oparta jest na eksperymentach z John Innes Centre we Wschodniej Anglii). Badaliśmy kilka modeli opartych na stochastycznych równań różniczkowych opisujących sygnał i wyróżniliśmy 2 reżimy (bazowy i impulsowy) oraz metody przejścia. W niektórych interakcjach bakteryjnych pojawia się trzeci system: strumień (duża fala wapnia) i opisowi tego zjawiska poświecony jest właśnie ten artykuł. Istnieje hipoteza, że ten strumień jest czymś w rodzaju protokółu komunikacji między bakteriami i roślinami. Wdrożyliśmy również mechaniczny algorytm automatycznie rozróżniający rodzaje interakcji symbiotycznej. Ten algorytm nie jest oparty na równaniach różniczkowych jak poprzednie modele, ale na geometrycznych właściwościach sygnału i bayesowskiej analizie widmowej. Automatyczna metoda rozpoznawania jest niezwykle ważna w interpretacji wyników, ponieważ eksperymentalne błędy, efekty stochastyczne i deterministyczne funkcje tła znacznie przewyższają siłę sygnału. Pytanie o kodowanie i dekodowanie za pomocą oscylacji wapniowych pozostaje nadal otwarte, ale nasza praca, jak również inne ostępy w dziedzinie bioinformatyki, może pomóc przetłumaczyć „protokół dyplomatyczny" pomiędzy Królestwem Roślin, a Królestwami Zwierząt, bądź Grzybów.

1. Wprowadzenie

Program został opracowany w celu automatycznej identyfikacji szczególnie wyjątkowego wzrostu stężenia wapnia w komórkach włosków korzeni po dodaniu czynnika Nod. Ten algorytm nie jest oparty na równaniach różniczkowych (Jarynowski 2010), ale geometrycznych właściwościach sygnału i bayesowskieh analizy widma. To bardzo ważne dla naukowców (algorytm rozpoznawania), ponieważ eksperymentalny szum, efekty stochastyczne i deterministyczne praw podstawowych w niebanalnej formie i wstępne rozumienie pochodzące z procesów biologicznych, często nie są wystarczająco dobrej jakości, aby umożliwić im standardowe analizy.

1.1. Założenia

Nasz program spogląda na geometryczne właściwości sygnału i próbuje rozpoznać, czy występują fala - strumień. Eksperymentatorzy definiujś strumień jako płynny wzrost współczynnika poziomu wapnia po dodaniu Nod, ale przed impulsowym obszarem systemu (gładkie wybrzuszenie na funkcji tła). Eksperymentatorzy mierzą poziom wapnia w dwóch miejscach: na końcówce (*tip*) i na komórce (*cell*). Nazywają to strumieniem tylko wtedy, gdy pojawia się równocześnie w obu miejacah.

1.2. Metody

W naszej analizie możemy zastosować dwie metody numeryczne do detekcji strumienia (*MinMax* i *Areas*). Dodatkowo metoda została poprawiona przez uwzględnienie lokalnych wariancji (w strumieniu nie ma impulsów, więc nasze wybrzuszenia powinny być łagodniejsze niż reszta sygnału). Prawdopodobieństwo warunkowe, wiedza *a priori* oraz spektralna analiza zostały użyte do wyeliminowania niepotrzebnych zmiennych.

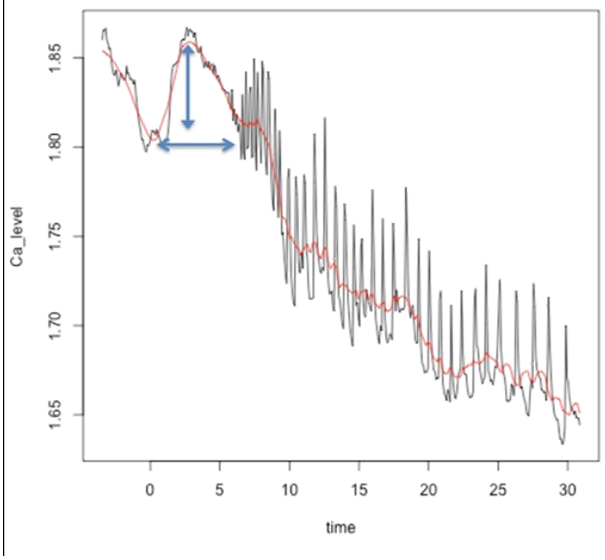
* Uniwersytet Jagielloński, Wydział Fizyki, Astronomii i Informatyki Stosowanej, ul. Reymonta 4, 30-058 Kraków

2. Algorytmy

Program jest napisany w środowisku R jako zestaw procedur. Wejściem dla tego programu są 3 wektory: czas, sygnał na końcówce (*tip*), sygnał na komórce (*cell*). Poza tym program działa na zasadzie quasi-prawdopodobieństwa pojawienia się strumienia w czasie. Program pozwala na automatyczne rozpoznanie pojawienia się strumienia.

2.1. MinMax

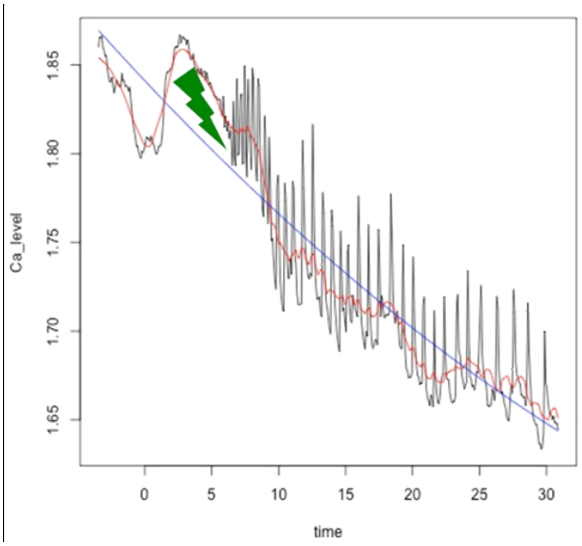
Badamy symetryczną średnią ruchomą, a dokładnie jej ekstrema. Odległość między dwoma lokalnymi minimami jest pomnożona przez odległość pomiędzy maksimum i średnią, tak że dwa minima stanowią podstawę do obliczenia prawdopodobieństwa (wcześniej należy unormować wymiar całego pudełka). Możemy zilustrować to na wykresie (Rys. 1), gdzie strzałki przedstawiają odległości opisane powyżej.



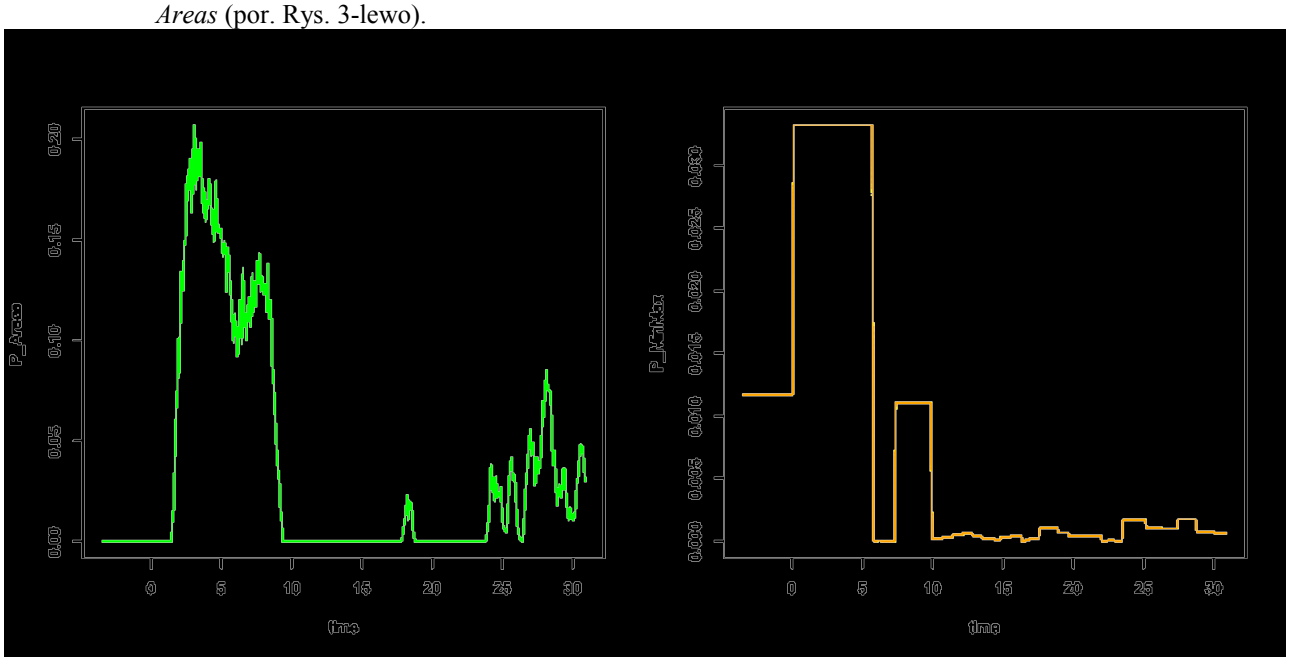
Rys. 1. Obserwacja z wyraźnie pozytywnym strumieniem (sygnał, MA(22)). Strzałki pokazują odległości stosowane w algorytmie *MinMax*.

2.2. Areas

Odkryliśmy, że najlepszym wyjaśnieniem tła (z uwzględnieniem złożoności) daje wielomian drugiego rzędu PF(2). Celem tej analizy jest znalezienie punktów przecięcia funkcyjnego dopasowania i MA(n) oraz obliczenie obszaru powyżej i poniżej MA(n). Metoda Monte Carlo została zastosowana w tych obliczeniach. Tak obliczone *Areas* są również znormalizowane przez wymiar pionowy wszystkich reakcji (odległość między minimalnymi wartościami a maksymalnymi).



Rys. 2. Obserwacja z wyraźnie pozytywnym strumieniem (sygnał, MA(22) i dopasowanie). Błyskawica pokazuje duży obszar między MA(22) i dopasowaniem wielomianowym: PF(2), które jest wskaźnikiem w algorytmie *Areas* (por. Rys. 3-lewo).



Rys. 3. Wykresy prawdopodobieństwa dla metody *Areas* (z lewej) i *MinMax* (po prawej).

Oba algorytmy (*MinMax* i *Areas*) nie biorą pod uwagę płynności sygnału. Aby dodać ten wątek użyjemy lokalnych wariancji. Przesuwanie odbywa się oddzielnie dla obu algorytmów. Po tym lokalnym zróżnicowaniu znormalizowanym przez globalną wariancję, sygnał został przesunięty w każdym wskazanym przedziale czasowym. Odchylenie zostało obliczone po detrendowieniu sygnału z MA(*n*). Przesunięcie może się odbyć jako wariancja, bądź odchylenie standardowe. Wybraliśmy przypadek pośredni – pomiędzy dwoma wykładnikami 1 a 1/2. Mnożąc przez czynnik *MinMax*:

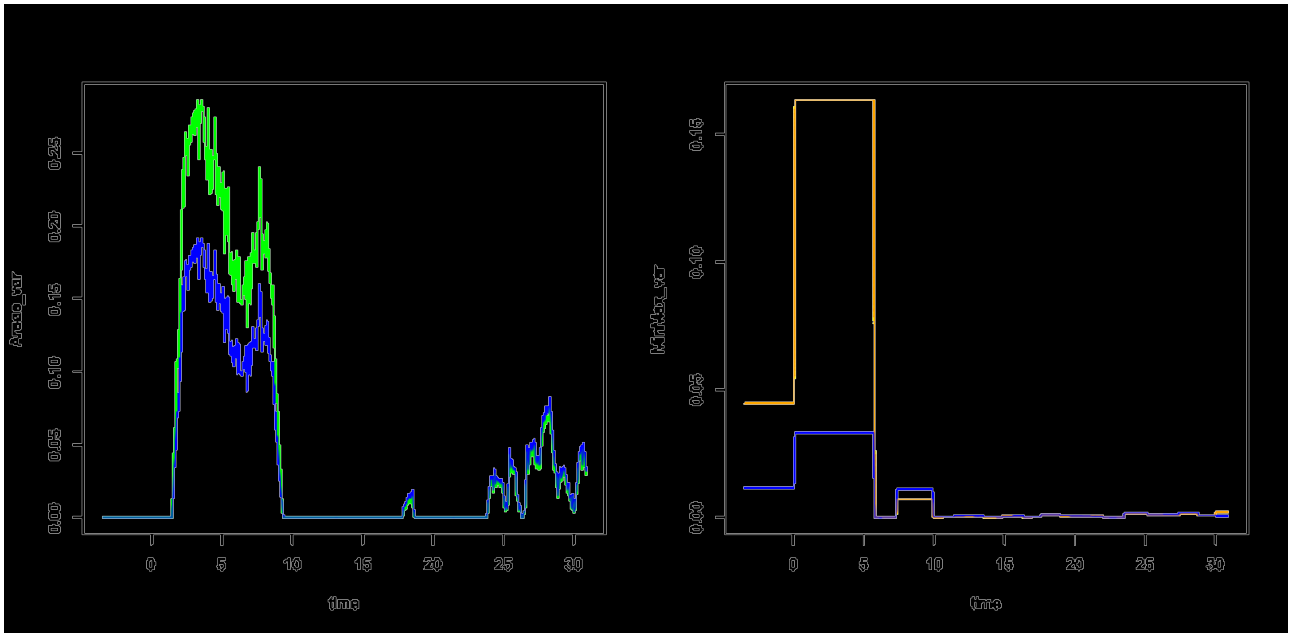
$$S_t^{MM} = Var^{0.75}(T)/Var^{0.75}(T_i) \qquad (1)$$

gdzie $t \in T_i$ (i-tego przedziału czasu między dwoma minimami wskazanymi przez *MinMax*) i cały szereg czasowy: $T = \cup T_i$.

Odpowiednio dla *Areas*:

$$S_t^{A} = Var^{0.75}(T)/Var^{0.75}(T_j) \qquad (1')$$

gdzie $t \in T_j$ (j-tego przedziału czasu między punktami przecięcia MA(n) z PF(2)) i $\cup T_j \in T$. Jeżeli $t \notin \cup T_j$ wtedy S_t^{A} nie istnieje. Aby zobaczyć, jak przesuwanie umożliwia lepsze rozpoznanie spójrzmy na Rys. 4.



Rys. 4. Wykresy quazi - prawdopodobieństwa dla metody *Areas* (z lewej) i *MinMax* (po prawej). Wyższe krzywe powstałe w wyniku przesunięcia przez wariancję. Musimy pamiętać, że operacja przeniesienia sprawia, że wynik nie może być rozumiany już jako prawdopodobieństwo w pełnym tego słowa znaczeniu.

3. Analiza probabilistyczna

W niniejszym rozdziale przedstawiono rozważania nad poprawieniem predykcji metodami bayesowskimi.

3.1. Dodawanie i mnożenie prawdopodobieństw

Załóżmy, że przesunięcie wartości uzyskane z algorytmów *MinMax* i *Areas* mogą być traktowane jako prawdopodobieństwa. Oznaczmy:
A - Istnieje strumień (wskazany przez geometryczne analizy)
A_{MM} - Istnieje strumień (wskazany przez przesunięte *MinMax*)
A_A - Istnieje strumień (wskazany przez przesunięte *Areas*)
$P(A_{MM})_t$ - Prawdopodobieństwo konieczności strumienia w punkcie *t* (zalecane przez przesunięte *MinMax*)
$P(A_A)_t$ - Prawdopodobieństwo strumienia w czasie *t* (zalecane przez przesunięcie *Areas*)
Pod względem logiki jeśli A_{MM} lub A_A będzie prawdziwe ($A_{MM} \vee A_A = A$) to zdarzenie A jest też prawdziwe. Możemy napisać równanie:

$$P(A_{MM} \vee A_A)_t = P(A_{MM})_t + P(A_A)_t - P(A_{MM} \wedge A_A)_t \quad (2)$$

Aby uprościć obliczenia przypuśćmy, że oba algorytmy nie znajdują strumienia w tym samym czasie, wtedy prawdopodobieństwo koniunkcji jest niewielkie { $P(A_{MM} \wedge A_A)_t << 1$}. Przy tym założeniu (zazwyczaj algorytmy znajdują strumień w różnym czasie) możemy $P(A_{MM} \wedge A_A)_t$ przyjąć za zero i (2) przyjmuje formę:

$$P(A)_t = P(A_{MM})_t + P(A_A)_t \quad (2')$$

Musimy pamiętać: poziom wapnia jest mierzony w dwóch miejscach: komórka (*cell*) i końcówka (*tip*) oraz dowody strumienia muszą być widoczne w obu miejscach pomiarowych. Aby odróżnić różne obserwacje przyjmijmy notację, np. A^C - Istnieje strumień w komórce; A^T - Jest strumień na końcówce. Zakładając, że A^C i A^T są niezależne, dostajemy:

$$P(A^C \wedge A^T)_t = P(A^T)_t \cdot P(A^C)_t \quad (3)$$

Równanie nie uwzględnia warunkowości, o czym w następnym punkcie {gdzie $P(A)_t = P(A^C \wedge A^T)_t$}.

3.2. Bayesowska analiza prawdopodobieństwa

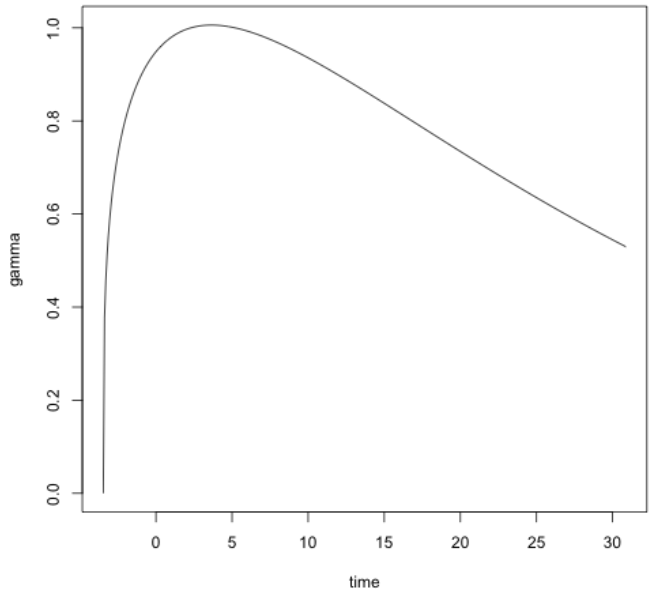
Aby lepiej zrozumieć proces, to strumień może się tylko zacząć kilka minut po fizycznym zainfekowaniu czynnika Nod. To daje nam *aprioryczną* wiedzę, która może być wykorzystana do opracowania prawdopodobieństwa. Przypomnijmy symbole i dodajmy nowe:
A - Istnieje strumień (wskazany przez geometryczne analizy)
B - Rzeczywiście istnieje strumień (zakładamy, że prawdopodobieństwo jest równe 1, bo nie analizujemy ukrytych strumieni)
Równanie Bayesa na prawdopodobieństwo warunkowe mówi:

$$P(A \setminus B) = \frac{P(B \setminus A) \cdot P(A)}{P(B)} \tag{4}$$

Warunkowe prawdopodobieństwa można rozumieć jako:
P(B) - zakładamy że jest 1, ponieważ nie analizujemy ukrytych strumieni
P(A) - prawdopodobieństwo wykrycia strumienia (wiedza *a priori* na temat czasu, kiedy strumień powinien się pojawić)
P(A\B) – prawdopodobieństwo występowania strumienia jeśli mamy przewidywania, że się pojawił: prawdopodobieństwo *a posteriori*
Teraz możemy przepisać (4) dla najważniejszych wartości prawdopodobieństwa *aposteriori* w funkcji czasu.
P(B\A) - prawdopodobieństwo konieczności strumienia uwarunkowane własnościami geometrycznymi (obliczone w ppkt 3.1)

$$P(A\setminus B)_t = P(B\setminus A)_t \cdot P(A)_t \tag{4'}$$

Nasze prawdopodobieństwa *a priori* $P(A)_t$ można opisać przesuniętą funkcją gęstości gamma. Parametry tej funkcji szacuje się, aby uzyskać jej maksimum w punkcie czasu, gdzie zwykle pojawia się strumień. Funkcja gęstości Gamma musiała zostać przesunięta, ponieważ jest określona tylko na wartościach nieujemnych (czas w naszym przypadku rozpoczyna się od wartości ujemnych, bo 0 oznacza, że czynnik Nod zaczyna mieć wpływ). To musiało być znormalizowane (aby uzyskać 1 dla maksymalnego punktu). Po tych manipulacjach prawdopodobieństwo *a priori* może być wykorzystane (4'). Przykład prawdopodobieństwa *a priori* jest pokazany na Rys.5. Prawdopodobieństwo *a priori* jest obliczane dla wszystkich objawów reakcji (zarówno na komórkę - *cell* i końcówkę - *tip*) .



Rys. 5. Funkcja *a priori* prawdopodobieństwa (zaczerpnięta z przesuniętej funkcji gęstości gamma). Maksymalna wartość prawdopodobieństwa jest usytuowana w jednym punkcie. .

3.3. Bayesowska redukcja parametrów

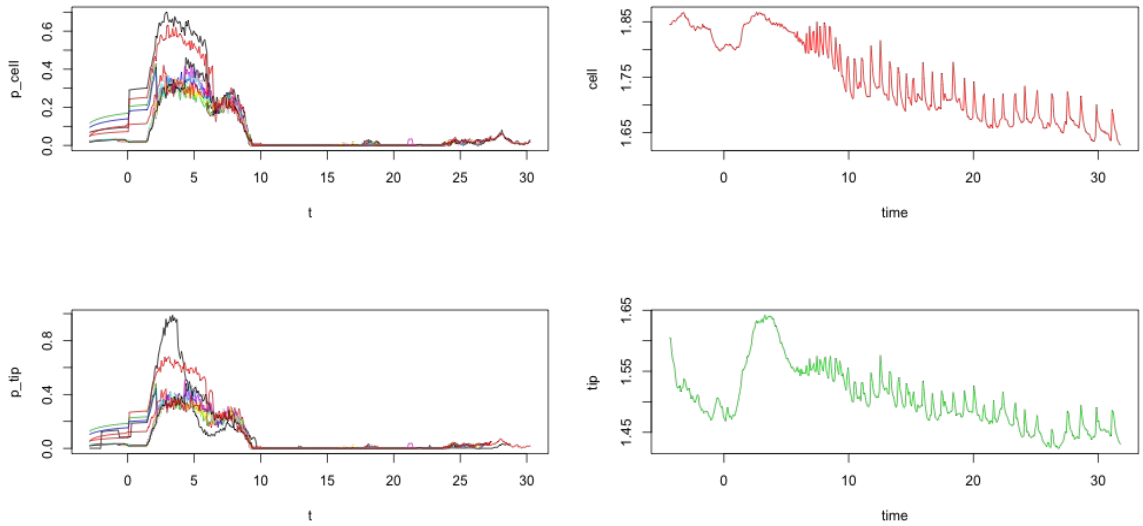
Predykcja naszych algorytmów zależy od parametrów stosowanych wewnątrz procedur (Bretthorst 1988). Załóżmy, że n jest "uciążliwym parametrem", którego nie trzeba znać. Rozszerzając nasze prawdopodobieństwo *a posteriori* z parametrem n, możemy zastosować twierdzenie Bayesa:

$$P(A,n \setminus B) = P(A,n)\frac{P(n \setminus A)}{P(n)} \tag{5}$$

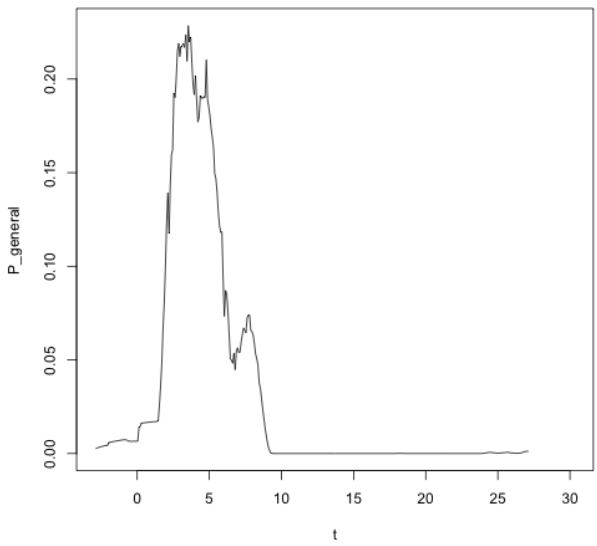
i scałkować po n,

$$P(A \setminus B)_t = \int dn P(A,n \setminus B)_t \tag{5'}$$

W naszym przypadku dokonaliśmy dyskretnego całkowania dla różnych n. Doświadczenie grupy naukowców z John Innes Centre wskazało za najbardziej adekwatne n=22 (Hazledine 2008) i liczymy średnie prawdopodobieństwo z wartości około tej liczby.



Rys. 6. Wykresy prawdopodobieństwa dla różnych n (z lewej) i sygnału (po prawej).



Rys. 7. Wykresy prawdopodobieństwa. Jest to produkt końcowy naszego programu. Poziom prawdopodobieństwa (żeby pozytywnie odpowiedzieć na pytanie, czy był strumień) musimy ustawić na 0,1. Na wykresie punkt maksymalny jest powyżej 0,2 – pozytywny przypadek, gdzie występuje strumień..

4. Dyskusja wyników i podsumowanie

Istniejący model opisujący dwie fazy: szumu i impulsów potrzebował rozwinięcia (Jarynowski 2010). Następnie spójrzmy na wady tego modelu (strumień będzie jednym z nich).
Doświadczenie w analizowaniu danych wapnia pomaga nam ustalić próg. Jeśli potraktujemy problem odwrotnie: możemy przebadać przypadki, które znamy i są pozytywne lub negatywne "na oko" i oszacować wartość progu. Proces ten daje na koniec poziom P=0,1 i to tylko sugestia dla eksperymentatorów (został wybrany arbitralnie). W przypadku niektórych nietypowych zestawów musieliśmy zmienić zasady (P=0,1). Dlatego należy sprawdzić:

- czy istnieje drugi szczyt przynajmniej wielkości połowy pierwszego i powyżej 0,075 (3/4 progu), to teraz próg wynosi 0,2
- jeśli istnieje trzeci lub dalszy szczyt zdecydowaliśmy się powiedzieć, że dane są zbyt zaszumione, aby ustalić czy to przypadek negatywny czy pozytywny
- jeśli szczyt jest węższy niż 3-4 minuty nie zaakceptujemy go.

Dzięki tym przepisom można ustrzec się złych wniosków. Badania szerokości szczytu można uniknąć przez dodanie kilku parametrów MA, ale to bardzo zwiększa złożoność obliczeniową, więc postanowiliśmy zrezygnować z tej opcji. Te dodatkowe założenia stosują się do tylko kilku procent odczytów i dotyczą tylko tych bardzo trudnych do analizy, ale są ważnym uzupełnieniem w zakresie statystycznej analizy ogromnej ilości danych.

5. Podziękowania

Ta praca nie powstałaby bez owocnych dyskusji z Richardem Morrisem (dyrektorem Zakładu Biologii Obliczeniowej i Systemowej, John Innes Centre w Norwich) jak również bez współpracy z doświadczalnikami również z John Innes Centre, a zwłaszcza z Giulią Morieri. Badania były również częścią międzynarodowego projektu (Leonardo da Vinci) obsługiwanego przez Politechnikę Wrocławską, a finansowane ze źródeł BBSRC (Biotechnology and Biological Sciences Research Council, UK).

Bibliografia

Hazledine, S. et al (2008) Differential and chaotic calcium signatures in the symbiosis signaling pathway of legumes. Proceedings of the National Academy of Sciences USA 105 (28) 9823-8

Bretthorst, G. L. (1988) Bayesian Spectrum Analysis and Parameter Estimation, Lecture Notes in Statistics, Springer Verrlag

Hazledine, S. et al (2009) Nonlinear time series analysis of nodulation factor induced calcium oscillations: evidence for deterministic chaos? PLoS One 4 (8) e6637

Jarynowski, A. (2010), Transmission patterns in changing states (spikes and noise) in plant's „neuronal" system. Stochastic Models in Neuroscience, CIRM, Marseille, Francja (notatki konferencyjne)