

SINTEZA LUCRARI

Titlu faza:

Analiza secventelor de codare din genomul bacterian prin modele autoregresive de ordin superior

Activitatile prevazute sunt:

- 1.1. Programare MATLAB verificare serii de date provenite din genomul bacterian al unor cazuri model
- 2.1. Analiza spectrala si identificarea modelului autoregresiv optim pentru specii bacteriene selectate; redactare articol pentru revista cotate ISI
- 2.2. Participare la The 7th European Biophysics Congress, Italia
Participare la The XX-th International Symposium on Bioelectrochemistry and Bioenergetics, Sibiu

Rezultate livrate pe etapa

1. Articole in revista indexata ISI.
 - **Microbial genome as a fluctuating system: Distribution and correlation of coding sequence lengths**
Vasile V. Morariu
REV. CHIM. (Bucuresti), 59, nr **11**, 1201-04, (2008)
 - **Autoregressive modeling of coding sequence lengths in bacterial genome**
Vasile V. Morariu, Luiza Buimaga-Iarinca
Trimisa la *Fluctuation and Noise Letters* ([arXiv:0907.1159](https://arxiv.org/abs/0907.1159)), in evaluare de catre referenti
2. Program editat in MATLAB pentru analiza structurii lungimii secventelor de codare din genomul bacterian: extragerea tendintei, efectuarea si medierea spectrelor, modelarea cu procese autoregresive de ordin superior.
3. Analiza cu programul elaborat mai sus, a genomului unor bacterii.

OBIECTIVELE ETAPEI:

1. Modelarea secventelor de codare din genomul bacterian prin modele autoregresive de ordin superior;
2. Cercetari privind organizarea si modelarea genomului bacterian la diverse specii.

SCOPUL ETAPEI:

1. Identificarea gradului optim al modelului autoregresiv care modeleaza seriile de date formate din lungimile secventelor de codare din genomul bacterian.
2. Determinarea, prin analiza parametrilor obtinuti, a diferentelor existente intre diferite specii bacteriene la nivel de codare a informatiei.

1. REZUMATUL ARTICOLELOR PUBLICATE:

1.1. *Microbial genome as a fluctuating system: Distribution and correlation of coding sequence lengths*

Genomul bacterian consta în principal din secvențe coding (90%), secvențele noncoding gasindu-se într-un procent mult mai mic. Lungimea acestor secvențe coding variază de la câteva sute până la câteva mii de perechi de baze. Valoarea medie a lungimii acestor secvențe este disponibilă în secțiunea de proteomică a bazei de date EMBL-EBI.

Distribuția lungimilor secvențelor coding a fost analizată prin densitatea de probabilitate a lungimilor. Proprietățile de corelare ale seriilor au fost investigate prin Analiza Fluctuațiilor Detrendate (DFA).

Au fost analizate diferite tipuri de distribuții, și am arătat că datele se potrivește cel mai bine cu o exponențială. Aceasta fitare este valabilă pentru cele mai multe genomuri, dar sunt și genomuri care prezintă deviații de la această lege. Credem că motivul pentru care apar aceste deviații este că seriile de lungimi de secvențe coding nu sunt uniforme pe o scară long-range.

1.2. *Autoregressive modeling of coding sequence lengths in bacterial genome*

Fiecare specie de bacterii sau archaea se caracterizează printr-o serie de secvențe de codare n fiecare având o anumită lungime lk , unde $k = 1 \dots n$. Lungimea este exprimată în număr de baze.

Analiza matematică a acestor serii a condus la doi parametri: α - exponent de corelare și ϕ - factorul de interacțiune. Aceștia au fost analizați în vederea explorării comparative a organizării în genomul unor bacterii folosite des în studii exploratorii.

Am analizat mai multe periodograme mediate ale acestor serii și am descoperit că organizarea CDS poate fi bine descrisă de modelul autoregresiv de ordin 1, care implică interacțiunea între termenii vecini. Parametru ϕ al procesului autoregresiv pare să fie mai sensibil decât exponentul de corelare.

2. PROGRAM DE MODELARE EDITAT ÎN MATLAB

Scopul programului realARp este de a estima modelul autoregresiv al unei serii de timp. Parametrii de intrare ai programului în ordinea în care apar la începutul programului (după instrucțiunea globală):

- *Numele fișierului* care conține valorile seriei de timp.
- *gradmin* și *gradmax* reprezintă gradul minim, respectiv maxim, al tendințelor polinomiale care sunt extrase din seria de timp pentru a obține spectrul pe care îl modelăm cu un proces AR(p)
- *Mvect* conține valorile intervalului de mediere al spectrului de putere
- *pmax* este ordinul procesului AR(p) cu care încercăm să modelăm seria de timp.
- *Ftaiat* ≤ 0.5 este frecvența de tăiere a spectrului, adică modelul AR(p) este obținut prin fitarea spectrului de putere al seriei de timp până la frecvența *ftaiat*. Dacă *ftaiat* = 0.5, atunci fitarea se face pe întregul spectru.

- *flagtaiat* este un indicator pentru alegerea modului în care este tratată porțiunea din spectru de putere cu frecvențe mai mari decât *ftaiat*.
- *S* este numărul de spectre de putere rarite aleator care sunt modelate cu procese $AR(p)$. Dacă $S = 0$, atunci modelarea se efectuează numai pe spectrul complet.

Materialele care conțin rezultatele modelări:

- Graficul *Extragerea tendinței* conține seria de timp originală și tendințele polinomiale cu gradele cuprinse între *gradmin* și *gradmax*.
- Graficul *Componenta aleatoare* conține valorile părții aleatoare a seriei de timp obținută prin scăderea tendinței polinomiale de gard *gardmax* din valorile seriei de timp originale.
- Graficele *Modelul $AR(p)$* conțin spectrele modelelor autoregresive obținute pentru intervalul de mediere specificat în titlul graficelor.
- În fișierul *rezultate.txt* și în fereastra de comandă sunt afișate sub formă tabelară rezultatele numerice corespunzătoare graficelor *Model $AR(p)$* . Fiecărei valori a intervalului de mediere îi corespunde un tabel separat. În plus la început este afișată și valoarea lungimii seriei de timp.
- Fișierul *paramARp.mat* conține rezultatele numerice ale modelării necesare pentru a fi transmise celorlalte programe.

3. ANALIZA UNOR GENOMURI MICROBIENE SI MODELAREA LOR CU PROCESE AUTOREGRESIVE DE ORDIN SUPERIOR

Fiecare moleculă de ADN se compune din două helixuri, *plus* și *minus* sau *dominant* și *recesiv*. Am investigat atât genomul ca întreg, cât și helixurile luate separat, notate $I(+/-)$, $I(+)$ și $I(-)$. Datele au fost extrase de pe site-ul web al Laboratorului European de Biologie Moleculară (EMBL) prin utilizarea unui program creat de noi. Un exemplu de genom bacterian este reprezentat în figura 1.

Aceste serii au fost investigate prin analiza fluctuațiilor detrendate de unde rezulta α - exponentul de corelare și modelarea cu model autoregresiv a spectrelor de putere, obținându-se ϕ – factorul de interacțiune.

Lista completă a speciilor analizate este: *Archaeoglobus Fulgidus*, *Bacillus subtilis*, *Bacillus Cereus ATCC14*, *Bacillus Halonduras*, *Escherichia Coli O157H7*, *Escherichia Coli Sakai*, *Escherichia Coli K12*, *Escherichia Coli UTI*, *Escherichia Coli W311*, *Escherichia Coli APEC*, *Hemophyliae Influentae 86*, *Hemophyliae Influentae ATC*, *Hemophyliae Influentae Pitt EE*, *Hemophyliae Influentae. Pitt GG*, *Helicobacter Pylori 266*, *Helicobacter Pylori HP*, *Helicobacter Pylori J99*.

2.1. Etape de lucru

Prin analiza fluctuațiilor detrendate am demonstrat că aceste serii prezintă corelare de distant scurtă.

Următorul pas constă din calculul periodogramelor și analiza acestora atât prin model autoregresiv de ordin unu, cât și prin model autoregresiv de ordin superior. În cele din urmă am comparat rezultatele obținute prin aceste tipuri de analize.

- DFA arată doar dacă în genom există sau nu corelare.

- Modelul autoregresiv demonstrează modul în care se realizează corelarea. Aceasta arată că termenii seriei formate din lungimile secvențelor coding interacționează și calculează taria de interacțiune.

Sucesiunea de operațiuni poate fi rezumată astfel:

1. Seria este detrendată prin diferențierea unui polinom cu gradul potrivit;
2. Se aplică transformata Fourier discretă;
3. Termenii periodogramei se mediază între 1-21 termeni;
4. Se ajustează spectrul cu un model AR(1) sau AR(p). Rezultate sunt parametrii factorul de interacțiune ϕ și dispersia σ .
5. Aceeași procedură a fost aplicată și pentru modelarea cu un model autoregresiv de ordin 4 notat AR(4).
6. Această procedură a fost aplicată pentru serii $I(+/-)$, $I(+)$ și $L(-)$.

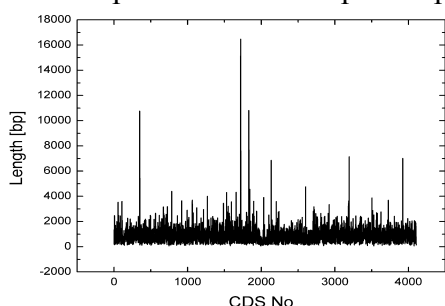


Fig.1 Seria de lungimi ale secvențelor coding pentru *Bacillus subtilis*. Este reprezentat întreg genomul, $I(+/-)$.

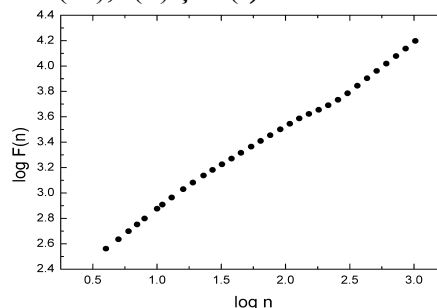


Fig.2 Analiza fluctuațiilor detrendate (DFA) pentru lungimile secvențelor de codare aparținând helixului plus al *Bacillus subtilis*.

2.2. Discutii

În fig. 3 și 4 este prezentat un exemplu de periodogramă mediata și nemediata pentru *Bacillus subtilis*. Se poate observa că periodograma nemediata permite fitarea liniară (fig 3.), ceea ce conduce la interpretarea greșită că este guvernată de o lege de putere (corelare de distanță lungă).

Forma reală a periodogramei este evidențiată după mediere (fig. 4), aceasta fiind mai bine fitată cu un model AR(1).

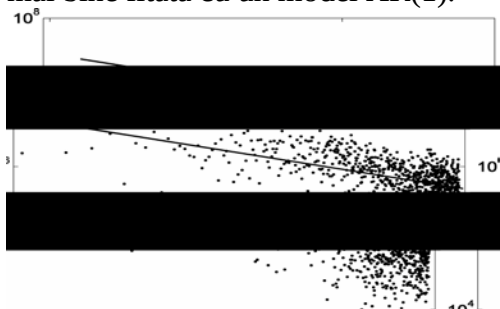


Fig.3 Periodograma nemediata pentru lungimile secvențelor coding pentru helixul plus (sau dominant) al *Bacillus subtilis*, fitată liniar

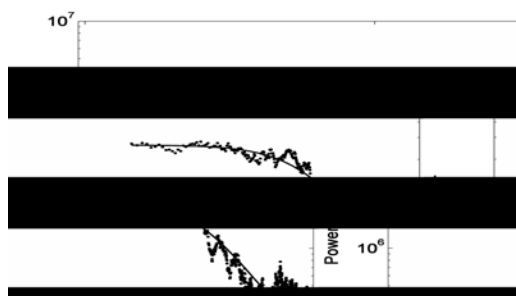


Fig.4 Periodograma mediata pentru lungimile secvențelor coding pentru helixul plus (sau dominant) al *Bacillus subtilis*, fitată cu model autoregresiv de ordin 1.

În fig. 5, 6 și 7 este ilustrată diferența între modelarea periodogramei aparținând seriei de lungimi ale secvențelor coding în genomul bacterian cu AR(4) și AR(9) față de AR(1).

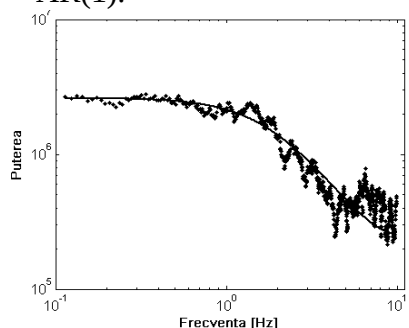


Fig. 5. Modelarea cu AR(1) a seriei de lungimi ale secvențelor coding pentru *B. subtilis* pe helixul dominant. Periodograma este mediata la 21 termeni.

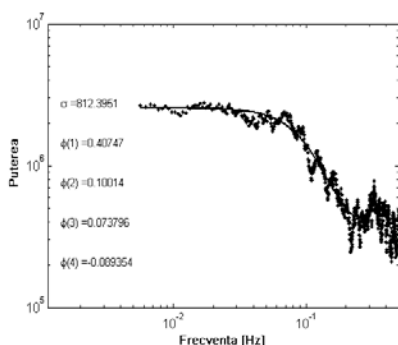


Fig. 6. Modelarea cu AR(4) a seriei de lungimi ale secvențelor coding pentru *B. subtilis* pe helixul dominant. Periodograma este mediata la 21 termeni.

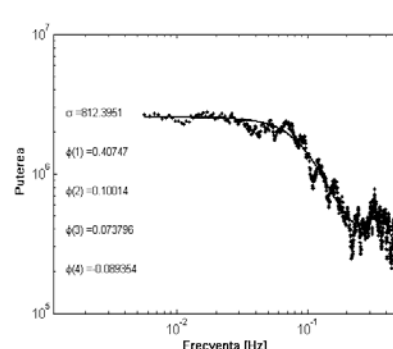


Fig. 7. Modelarea cu AR(9) a seriei de lungimi ale secvențelor coding pentru *B. subtilis* pe helixul dominant. Periodograma este mediata la 21 termeni.

Se observă că deși modelul de ordin superior se fitează mai bine pe periodograma, modelul autoregresiv de ordin unu este totuși suficient pentru modelarea acestor tipuri de serii.

În cazul unor bacterii și archaea prezentate în tabelul 2, acest parametru variază între 0.52 pentru *Bacillus subtilis* și aproape 0 pentru unele tulpini de *Haemophilus influenzae* și *Helicobacter pilori*. Aceasta înseamnă că taria interacțiunii dintre termenii seriilor formate din lungimile secvențelor coding poate să varieze mult, de la o valoare medie în jos, până la o interacțiune aproape nulă. Mai mult, această interacțiune este sensibilă la diferențele dintre helixul dominant și cel recesiv.

Se poate observa și diferența dintre specii și/sau dintre tulpinile aceleiași specii. În plus, exponentul de corelare are o valoare mică, ceea ce indică o corelare slabă între termenii seriei.

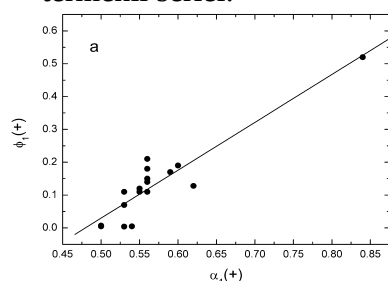


Fig. 8.a Relația dintre exponentul de corelare α și factorul de interacțiune ϕ pe helixul + pentru toate genomurile investigate

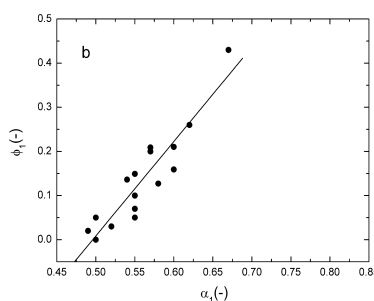


Fig. 8.b Relația dintre exponentul de corelare α și factorul de interacțiune ϕ pe helixul - pentru toate genomurile investigate

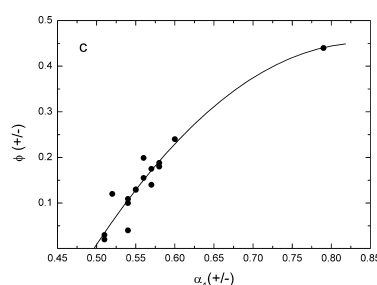


Fig. 8.c Relația dintre exponentul de corelare α și factorul de interacțiune ϕ pentru toate genomurile investigate

Este evident că există o relație între exponentul de corelare α și factorul de interacțiune ϕ . Aceasta este ilustrată în figura 8 pentru trei tipuri de serii. Se poate

observa o dependenta liniara intre α si ϕ pentru $l(+)$ si $l(-)$, in vreme ce pentru intreg genomul aceasta nu mai este liniara.

Se pare ca parametrul ϕ , rezultat din analiza cu modelul autoregresiv este mult mai sensibil decat exponentul spectral de corelare.

De asemenea, speram ca am dezvoltat un model matematic care sa poata fi folosit si la analiza altor sisteme naturale.

CONCLUZII:

- S-a efectuat analiza fluctuatiilor detrendate si analiza autoregresiva pe periodogramele mediate ale seriilor date de lungimile secventelor coding pentru diferite specii de bacteria si archaea.
- Parametrul obtinut din analiza fluctuatiilor detrendate nu este foarte sensibil. In plus, forma reprezentarii grafice a DFA indica o corelare de distant scurta.
- Analiza cu modelul autoregresiv de ordin 1 este suficienta pentru modelarea spectrelor mediate (periodograme).
- Modelele autoregresive de ordin superior fiteaza mult mai bine periodograma, dar informatiile oferite sunt similare celor obtinute prin fitarea cu AR(1).
- In analiza secventelor coding in genom, se impune separarea datelor pe cele doua helixuri.
- Dependenta dintre exponentul spectral de corelare si factorul de interactiune este liniara pentru cele doua helixuri (dominant si recesiv).
- Cand investigam intreg genomul, dependenta nu mai este liniara, cele doua helixuri interfera si actioneaza unul asupra celuilalt similar unor zgomete.