UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Modele de învățare automată în proteomică

Rezumatul Tezei de Doctorat

Student doctorand: Silvana Albert
Conducător de doctorat: Prof. Dr. Gabriela Czibula

2019

Cuvinte cheie: învățare automată, proteomică, pliere proteică, clustering, self organizing maps, învățare nesupravegheată, data mining, software open source

Cuprinsul rezumatului

Cuprinsul tezei

Lista publicațiilor

Toate clasamentele sunt listate în conformitate cu clasificarea jurnalelor din 2014 [1] și conferințe [2] în informatică

1. [AC19] **Silvana Albert**, Gabriela Czibula. $ProteinA$: An approach for analyzing and visualizing protein conformational transitions using fuzzy and hard clustering techniques, 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019), va fi publicat. (**ISI Proceedings**)

   Rang B, 4 puncte.

2. [ATC18] **Silvana Albert**, Gabriela Czibula, Mihai Teletin. Analysing Protein Data Using Unsupervised Learning Techniques. *International Journal of Innovative Computing, Information and Control*, Volumul 14, Numarul 3, Iunie 2018 14(3):861–880, 2018. (**indexed Scopus**)

   Rang B, 4 puncte.

3. [TCAB18] Mihai Teletin, Gabriela Czibula, **Silvana Albert**, Maria-Iuliana Bocicor Using unsupervised learning methods for enhancing protein structure insight. *International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2018)*, Volumul 126, 19-28, 2018 (**ISI Proceedings**)

   Rang B, 2 puncte.

4. [TCB+18] Mihai Teletin, Gabriela Czibula, Maria-Iuliana Bocicor, **Silvana Albert**, and Alessandro Pandini. *Deep autoencoders for additional insight into protein dynamics*. International Conference on Artificial Neural Networks (ICANN), Rhodes, Greece, LNCS, Volumul 11140, pp. 78-89, 2018.

   Rang B, 1.33 puncte.

5. [ACT18] **Silvana Albert**, Gabriela Czibula, Mihai Teletin. Analyzing the impact of protein representation on mining structural patterns from protein data. *IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI 2018)* , Volumul 12, 533-538, 2018 (**ISI Proceedings**)

[1] http://informatica-universitaria.ro/getpfile/16/CSafisat2.pdf;http://hfpop.ro/standarde/doctorat/2014-jurnale.pdf

[2] http://informatica-universitaria.ro/getpfile/16/CORE2013_Exported.xlsx;\http://hfpop.ro/standarde/doctorat/2014-conferinte.xlsx

Rang C, 2 puncte.

6. [Alb17b] **Silvana Albert**. Visualizing mutation occurrence using Big Data. *Proceeding Collection on the workshop NETTAB 2017 "Methods, tools and platforms for Personalized Medicine in the Big Data Era"*, 2017. (**indexat DBLP**)

   Rang D, 1 punct.

7. [Alb17a] **Silvana Albert**. A Big Data Approach in Mutation Analysis and Prediction. *Studia Universitatis Babes-Bolyai Series Informatica*, 62(1):75–89, 2017. (**indexat Mathematical Reviews**)

   Rang D, 1 punct.

8. [BPC+17] Maria Iuliana Bocicor, Alessandro Pandini, Gabriela Czibula, **Silvana Albert**, Mihai Teletin. Using Computational Intelligence Models for Additional Insight into Protein Structure. *Studia Universitatis Babes-Bolyai Series Informatica*, 62(1):107–119, 2017. (**indexat Mathematical Reviews**) **Rang D, 0 puncte.**

Scorul publicațiilor: **15.33 puncte**.

Introducere

Această teză de doctorat este rezultatul cercetărilor mele efectuate începând cu 2016 sub supravegherea Prof. Dr. Gabriela Czibula. Ea constă în principal în dezvoltarea de modele de învățare automată și aplicarea lor pe date biologice împreună cu analiza rezultatelor cu scopul mai mare de a descoperi motive ascunse în datele genomice și proteomice. În această teză folosim diverse tehnici de învățare automată pentru a obține o mai bună cunoaștere a proceselor biologice. Spațiul de date proteice, în special, va fi explorat cu ajutorul diferiților algoritmi, cum ar fi *clustering*, *self-organizing maps*, rețele neuronale de tip *Long-Short Term Memory*, Analiza Componentelor Principale și *Autoencoders*. Descoperirea tiparelor proteice ajută la o mai bună înțelegere a modului în care proteinele se schimbă și evoluează. Este deja cunoscut faptul că forma proteinei dictează funcția biologică, astfel încât orice anomalie în timpul procesului de pliere poate duce la o proteină malformată, defectă, care, în timp, poate provoca diverse boli sau chiar moartea organismului. De aceea, trebuie să investigăm metode de calcul și să dezvoltăm soluții care să îi ajute pe cercetători să înțeleagă mai bine lumea misterioasă a plierii proteinelor. Interesul, pasiunea și fascinația crescută a autorului pentru procesele biologice care creează și asigură viața reprezintă motivația alegerii acestei teme. Atât importanța, cât și complexitatea problemei ne motivează să explorăm utilitatea modelelor și metodelor de învățare automată pentru analiza și detectarea modificărilor conformaționale ale proteinelor. Munca noastră conține diferite aplicații ale algoritmilor de învățare automată pe date biologice, împreună cu analiză rezultatelor, cu scopul mai mare de a descoperi tipare și motive ascunse în datele proteice. Ne orientăm studiile noastre și asupra enzimelor care au proprietăți interesante atunci când vine vorba despre biodegradarea deșeurilor. Pe partea de învățare automată a existat, de asemenea, o creștere fără precedent a interesului și popularității. În timpul procesului de pliere proteică, proteina suferă modificări de la o conformație la alta și este influențată atât intern, de către structura inițială și compoziția sa, precum și extern, de către alți factori, cum ar fi temperatura și alte elemente din apropiere care interacționează cu aceasta. Înțelegerea dinamicii proteinei duce la crearea de medicamente mai precise, încetinește procesul de îmbătrânire și îmbunătățește calitatea vieții. Cu scopul principal de a înțelege importanța problemei de pliere a proteinelor și de a descoperi motive ascunse în datele proteice, am analizat tranzițiile conformațiilor proteice folosind instrumente de învățare nesupervizate, prin aplicarea diferitelor tipuri de algoritmi de clustering hard și fuzzy și compararea rezultatelor. Valorile RSA au fost studiate și relevanța lor a fost testată atunci când este vorba de predicția tranzițiilor interne ale proteinelor. Folosind aceste valori, am făcut mai multe studii de caz din domeniul învățării nesupravegheate. Pentru o mai bună vizualizare, am folosit Analiza Componentelor Principale și am putut vedea cum evoluează proteinele, prin valorile lor RSA, de la o conformație

la alta. S-a utilizat Deep Learning, în special rețelele de tip Long Short Term Memory pentru explorarea tranzițiilor conformaților proteice, utilizând o altă variabilă proteică: ”width of the binding cleft” care ajută la obținerea mai multor informații. Am introdus si două soluții software open source care permit experților din alte domenii să reproducă cu ușurință experimentele noastre ducând la o mai bună colaborare interdisciplinară. Abordările noastre au fost publicate în reviste și conferințe: [Alb17a, BPC+17, ATC18, TCAB18, ACT18, TCB+18, AC19, Alb17b].

Capitolul 1: Analiza datelor proteice folosind învățarea nesupravegheată

În acest capitol cercetăm utilitatea modelelor și metodelor de învățare automată pentru analiză și detectarea modificărilor conformaționale ale proteinelor. Prezentarea din acest capitol se bazează pe lucrările noastre originale [BPC+17], [ATC18] și [ACT18].

2.1 Un model teoretic pentru analiza tranzițiilor conformaționale proteice

Abordăm problema determinării tranzițiilor conformaționale proteice dintr-un alt unghi și derivăm pentru aceasta o formalizare diferită, pornind de la un set de date de peste 300 de proteine și conformațiile lor asociate. Obiectivul nostru este să testăm dacă tranzițiile conformaționale ale proteinelor oferă informații utile cu privire la structura lor tridimensională și dacă un model de *învățare nesupervisată* este capabil să surprindă acest tip de relații biologice între proteine. Modelul ales de învățare automată este un emph self organising map (SOM), deoarece este considerat un instrument puternic din domeniul *dată mining* care ajută la vizualizarea datelor cu multe dimensiuni. Folosind un experiment de *dată mining*, arătăm că informațiile obținute prin analizarea tranzițiilor conformaționale ale proteinelor sunt capabile să surprindă relațiile dintre proteinele înrudite, relații care sunt confirmate din perspectivă biologică.

2.2 Analiza datelor proteice folosind Self Organizing Maps și Analiza Componentelor Principale

Proteinele au roluri esențiale în procesele biologice ale organismelor vii, contribuind la menținerea mediilor celulare. Înțelegerea tranzițiilor conformaționale ale proteinelor poate ajuta la identificarea situațiilor în care pot apărea plieri incorecte sau mutații și, astfel, inhibând posibilului comportament necontrolat și nedorit. Similitudinea structurală dintre proteine este descoperită într-o manieră nesupervizată, folosind emph Self-Organizing Maps de tip hard și fuzzy, bazate pe tranzițiile conformaționale ale proteinelor. Propunem o metodă de modelare a proteinelor pe baza tranzițiilor conformaționale și examinăm, de asemenea, modul în care selecția unor caracteristici afectează performanța modelelor propuse.

2.3 Analiza impactului reprezentării proteinelor asupra modelelor structurale de *Data Mining* din datele proteice

Efectuăm un studiu despre modul în care reprezentările diferite ale proteinelor au impact asupra procesului de *mining* al modelelor relevante din datele proteice. Două reprezentări sunt utilizate pentru proteine, una folosind alfabetul structural și a doua folosind valorile RSA ale aminoacizilor din structura primară a proteinelor. Folosind aceste reprezentări, două studii de caz sunt realizate pentru a sublinia eficiența utilizării reprezentărilor proteice propuse pentru a învăța, intr-o manieră nesupervizată, tiparele structurale dintr-un set de date proteice. Valorile RSA par a fi foarte relevante pentru a reprezenta conformațiile proteinelor.

Capitolul 2 - Abordări de *clustering* pentru analiza și vizualizarea datelor proteice

În acest capitol folosim *clustering* ca metodă de clasificare nesupravegheată pentru a studia relevanța valorilor *RSA* pentru a analiza tranzițiile proteice interne. Oferim două abordări care implică clustering hard și soft iar apoi discutam rezultatele: [TCAB18], [AC19].

3.1 Metoda 1: Un model teoretic pentru îmbunătățirea cunoștințelor despre structura proteinelor

Dezvoltăm un studiu orientat spre investigarea modului în care evoluția tranzițiilor conformaționale ale proteinelor evoluează în timp, cu scopul de a lărgi cunoștințele în dinamica internă a proteinelor. Utilizăm clustering-ul ca tehnică de învățare nesupervizată pentru a determina importanța valorilor RSA în decodificarea tranzițiilor proteice. Pentru fiecare dintre proteine avem 10000 conformații disponibile, împreună cu valorile RSA asociate. Acestea au fost obținute prin simulări de dinamică moleculară, un proces care este considerat destul de costisitor din punct de vedere al resurselor necesare. Conformațiile sunt consecutive și pot fi privite ca un moment in timp din etapa de formare a proteinei. Ipoteza noastră este că schimbările pe care o proteină le suferă de la o conformare la următoarea sunt foarte mici (dacă există astfel de modificări) ca o deducție din perspectiva biologică, de unde deducem urmatoarea concluzie: conformațiile consecutive apropiate sunt foarte similare.

De aceea calculăm distanța euclidiană pentru a obține disimilitudinea dintre două tranziții conformaționale succesive. Seturile de date rămân fără etichete pe parcursul întregului proces, învățarea nesupravegheată fiind strategia noastră. Deoarece proteină suferă modificări conformaționale, anumite părți ale structurii sale sunt supuse unor modificări minore, care se reflectă în pozițiile reziduurilor de aminoacizi și, în consecință, în valorile RSA ale acestora. Astfel, conformațiile consecutive sunt destul de similare din perspectiva reprezentărilor lor considerate (valorile RSA). Acest lucru se reflectă și în rezultatele obținute prin clustering. Se observă, de asemenea, că proteinele sunt similare structural și, din această cauză, se anticipează un grad ridicat de similitudine în forma și dinamica lor reprezentată. Subliniem potențialul modelelor de clustering pentru modelarea corectă a tranzițiilor biologice, din conformațiile reprezentate prin valorile RSA. Din punct de vedere biologic, într-o tranziție dintre două conformații succesive, proteina poate rămâne neschimbată sau pot apărea ușoare modificări în anumite părți ale acesteia.

3.2 Metoda 2: Un model pentru analiza tranzițiilor conformaționale proteice folosind tehnici de clustering *fuzzy* și *hard*

Realizăm mai multe experimente pe două seturi de date proteice, cu scopul de a arăta empiric, folosind mijloacele de clustering *fuzzy c-means* și *Birch*, că valorile RSA se schimbă lent atunci când o proteină trece prin schimbări conformaționale. Scopul final este să demonstram că tranzițiile conformaționale consecutive sunt mai apropiate și proteina evoluează liniar. Cele două proteine utilizate în experimentele noastre sunt: 6EQE - o structură cristalină de înaltă rezoluție a unei hidrolaza degradantă din polietilenă tereftalată din Ideonella sakaiensis" și 4CG1 - o hidrolază degradantă din polietilen tereftalat termostabilă de la Thermobifida fusca" [WOT+14]. Ambele proteine sunt cercetate pentru rolurile lor în degradarea PET-ului și faptul că au multe în comun ne determină să anticipam rezultate similare atunci când aplicăm algoritmi nesupervizați. Notăm că operațiunea fuzzy nu se îmbunătățește așa cum era de așteptat. Acest lucru se datorează posibil faptului că datele de intrare nu sunt neaparat adecvate pentru a fi grupate in stări, având în vedere reprezentarea aleasă. In viitor, vom continua în această direcție, pentru identificarea altor reprezentări îmbunătățite, mai mult pentru perspectiva fuzzy, deoarece nu îmbunătățește eficacitatea procesului de clustering. Vor fi efectuate mai multe experimente în această direcție.

Capitolul 3 - Abordări de *Deep Learning* pentru analiza datelor proteice

Acest capitol este orientat în jurul Rețelelor Neuronale și modul în care acestea pot fi utilizate pentru captarea motivelor ascunse în datele proteice. Explorăm emph Deep Autoencoder Neural Networks și capacitatea lor de a captura aspecte legate de structura unei proteine. Secțiunea reprezintă o continuare a lucrărilor noastre prezentate în [TCB+18].

4.1 Folosirea Autoencoderelor pentru descoperirea dinamicii proteinelor

Deoarece abordările dinamicii moleculare sunt atât de costisitoare, reducerea dimensionalității datelor este, de asemenea, obiectivul nostru. *Denoising sparse autoencoders* sunt instruite pe fiecare set proteic de date cu scopul principal de a reduce dimensionalitatea seturilor de date și de a ajuta la vizualizare. Pentru validarea rezultatelor noastre, calculăm similitudini în setul de date inițial și, în final, vom calcula asemănările datelor bidimensionale emise de autoencoder.

4.2 Prezicerea lățimii ”binding cleft”-ului prin utilizarea rețelelor neurinale de tip *Long short-term memory networks*

Intuiția noastră este că dacă luăm în considerare informații suplimentare despre fiecare conformație proteică, vom avea o mai bună perspectivă asupra funcționării interne a unei proteine în timpul procesului de pliere. Deoarece folosim o reprezentare 2D a unui obiect 3D, pierdem informații valoroase despre proteine. Includerea altor valori care descriu fiecare conformație este încercarea noastră de a recâștiga o parte din informația pierdută. Acesta este motivul pentru care exploatăm o proprietate numită emph width of the binding cleft (cunoscută și sub numele de emph active-site cleft), care le caracterizează atât pe emph Cutinase cât și pe emph PETase. Studiile au fost făcute pentru a demonstra că, minimizarea valorii acestei proprietăți pentru că *PETase* să semene mai mult cu *Cutinase*, ar duce la o mai bună degradare a PET-ului [AAD+18]. Utilizăm o rețea neuronală de tip *Long short term memory network* pentru a afla informații importante despre tranziția unei proteine și, în principal,

modul în care fiecare conformație poate fi evaluată atunci când vine vorba despre starea ei deschisă/ închisă. Folosind mai multe împărțiri de date *train /test* în procesul de învățare și testare, încercăm să îmbunătățim performanța modelului pe datele noi, datorită faptului că mai multe modele sunt astfel antrenate. Cu toate acestea, suntem conștienți că abordarea de împărțire mulți-test are o limitare: fiecare dintre modelele de antrenare rămâne fix, pe măsură ce este evaluat în setul de test. După instruirea rețelei LSTM, suntem capabili să prezicem noi valori pentru proprietatea descrisă mai sus, pe baza reprezentarii prin unghiuri a conformațiilor proteice. Concluzionăm că LSTM-urile sunt adecvate pentru problema prezentată, iar modelul de calcul dezvoltat este capabil să prezică valorile viitoare pentru proprietatea investigată. Din perspectivă biologică, acest lucru este util, deoarece, făcând mici modificări la proteinele de tip emph PETase, pot să funcționeze mai bine când vine vorba de degradarea PET-ului [AAD+18]. Modele similare ar putea fi utilizate de oamenii de știință, deoarece modifică compoziția proteinelor, pentru a emula starea emph deschis / închis. Direcțiile viitoare includ o rețea LSTM de clasificare care ar putea determina cu ușurință dacă o nouă instanță (conformarea proteinei în cazul nostru) are starea deschisă sau închisă.

Capitolul 4 - Dezvoltarea de *Software*

În acest capitol introducem două soluții software care reprezintă munca noastră originală publicată în [Alb17a], [Alb17b] și [AC19]. Încercăm prin aceste aplicații să contribuim la comunitatea open-source și să permitem experților din diverse domenii să reproducă cu ușurință experimentele noastre ducând la o mai bună colaborare interdisciplinară.

5.1 ProteinA: O soluție software pentru vizualizarea tranzițiilor proteice

Software-ul prezentat în această secțiune a fost creat pentru a permite utilizatorilor să încerce diverse combinații sau parametri și să analizeze în mod independent rezultatele. A fost introdus pentru prima dată în lucrarea noastră originală [AC19]. Propunem aplicația *ProteinĂ* pentru captarea tranzițiilor conformaționale proteice prin clustering. Este o aplicație web care permite utilizatorilor să ruleze analize personalizate și să descarce rezultatele. O analiză de clustering durează aproximativ 5 minute, însă ideea din spatele software-ului este de a permite procesare mai complexă și de a oferi rezultatele când este gata. Soluția la care se rulează public la [Alb19][1]. Codul este disponibil pe Github at [Alb18b].[2] O altă opțiune pentru a o rula cu ușurință pe o mașină locală este accesarea imaginii de docker, publică la [Alb18a]. [3]

5.2 Software nou pentru vizualizarea mutațiilor genetice

Soluția prezentată ajută prin agregarea tuturor mutațiilor precedente corelate cu o serie de factori externi. Medicul este capabil să restrângă diagnosticul la un număr rezonabil de posibilități, pe baza cazurilor deja rezolvate. Acest lucru duce la luarea unei decizii în cunoștință de cauză pentru care dintre mutații trebuie testate. După ce a pus diagnosticul cu succes, specialistul îl va introduce în baza de date globală, în acest fel, ajutând viitorii medici. Ca dovadă a conceptului, aplicația demonstrează rolul uriaș pe care Big Dată îl

[1]Protein clustering online http://proteinclusters.online/proteins.

[2]Protein clustering web application https://github.com/albusilvana/proteinclusteringwebapp.

[3]Protein clustering docker image on Docker hub https://hub.docker.com/r/salbert/proteinclustering.

are în agregarea mutațiilor genetice și poate fi considerat un punct de plecare pentru soluții similare care urmăresc să inoveze genetica.

Bibliografie

[AAD+18] Harry P. Austin, Mark D. Allen, Bryon S. Donohoe, Nicholas A. Rorrer, Fiona L. Kearns, Rodrigo L. Silveira, Benjamin C. Pollard, Graham Dominick, Ramona Duman, Kamel El Omari, Vitaliy Mykhaylyk, Armin Wagner, William E. Michener, Antonella Amore, Munir S. Skaf, Michael F. Crowley, Alan W. Thorne, Christopher W. Johnson, H. Lee Woodcock, John E. McGeehan, and Gregg T. Beckham. Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences*, 115(19):E4350–E4357, April 2018.

[AC19] S. Albert and G. Czibula. $proteina$: An approach for analyzing and visualizing protein conformational transitions using fuzzy and hard clustering techniques. In *12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019)*. Springer, May 2019.

[ACT18] S. Albert, G. Czibula, and M. Teletin. Analyzing the impact of protein representation on mining structural patterns from protein data. In *IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI 2018)*, page To be published, 2018.

[Alb17a] S. Albert. A big data approach in mutation analysis and prediction. *Studia Universitatis Babes-Bolyai Series Informatica*, 62:75–89, 2017.

[Alb17b] S. Albert. Visualizing mutation occurrence using big data. In *ProceedingCollection on the workshop NETTAB 2017 Methods, tools platforms for PersonalizedMedicine in the Big Data Era, 2017*, 2017.

[Alb18a] Silvana Albert. Protein clustering docker image. https://hub.docker.com/r/salbert/proteinclustering, 2018.

[Alb18b] Silvana Albert. Protein clustering git repository. https://github.com/albusilvana/proteinclusteringwebapp, 2018.

[Alb19] Silvana Albert. Protein clustering analysis, 2019.

[ATC18] S. Albert, M. Teletin, and G. Czibula. Analysing protein data using unsupervised learning techniques. *International Journal of Innovative Computing, Information and Control*, page 861–880, 2018.

[BPC+17] M.I. Bocicor, A. Pandini, G. Czibula, S. Albert, and M. Teletin. Using computational intelligence models for additional insight into protein structure. *Studia Universitatis Babes-Bolyai Series Informatica*, 62:107–119, 2017.

[TCAB18] M. Teletin, G. Czibula, S. Albert, and I. Bocicor. Using un-supervised learning methods for enhancing protein structure insight. In *InternationalConference on Knowledge Based and Intelligent Information and EngineeringSystems*, page To be published, 2018.

[TCB+18] M. Teletin, G. Czibula, M.I. Bocicor, S. Albert, and A. Pandini. Deep autoencoders for additional insight into protein dynamics. In *22nd International Conference on Knowledge-Based and Intelligent Information Engineering Systems*, pages 79–89. Springer, 2018.

[WOT+14] Ren Wei, Thorsten Oeser, Johannes Then, Christina G. Föllner, Wolfgang Zimmermann, and Norbert Sträter. Structural and functional studies on a thermostable polyethylene terephthalate degrading hydrolase from thermobifida fusca. *Applied Microbiology and Biotechnology*, 98:7815–7823, 2014.