Modelarea si simularea activitatii pompelor de eflux bacteriene prin metode laser avansate (AMPLE)

Iradierea cu fascicule laser a medicamentelor, caracterizarea lor spectrala si stabilirea nivelului de activitate antibacteriana a lor

LSG/INFLPR

Rezultatele obtinute de unitatea coordonatoare se inscriu sub titlul generic: Studii de docking molecular pentru caracterizarea medicamentelor in vederea stabilirii activitatii de inhibare a pompelor de eflux (epi) – calculul structurii electronice si al dockingului molecular. In cadrul etapei s-a realizat un studiu de docking molecular pentru o proteina de efflux din *S. aureus*, NorA, si cativa compusi cu proprietati de inhibare a pompelor de eflux din clasa fenotiazinelor si chinazolinelor. Studiul as fost structurat in trei etape. In prima etapa s-au realizat modelele moleculelor de fenotiazine, chinazoline si a unor compusi naturali. In etapa a 2 s-a realizat un studiu de predictie a structurii proteinei NorA folosind programul I-Tasser si baza de date RCSB PDB (protein databank). In etapa a treia s-au realizat studii de docking molecular cu compusii chimici selectati si pompa de eflux NorA si s-a analizat capacitatea de predictie a activitatii EPI. Rezultatele studiului au fost redactate sub forma unui articol in pregatire pentru publicare (anexat raportului).

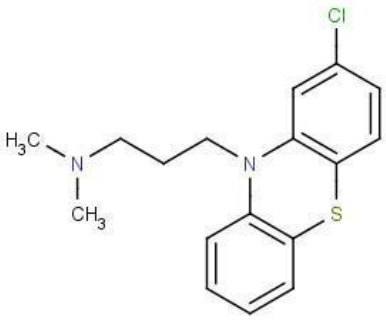
Un alt subiect l-a constituit analiza efectelor iradierii medicamentelor cu laseri in UV. Pentru investigarea efectului iradierii cu laser asupra medicamentelor au fost stabilite cai de reactie posibile si s-au realizat modelele de structura optimizate pentru compusi rezultati din iradiarea moleculelor de medicament. Caile de reactie sunt studiate folosind programul Gaussian.

Rezultate

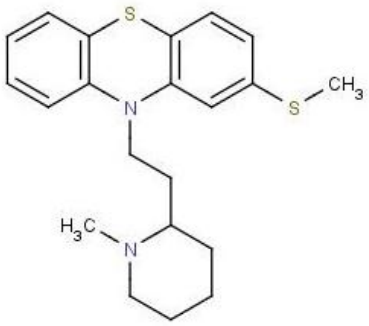
1. Calcul de structura electronica

Mai multi compusi chimici au fost studiati ca potentiali inhibitori ai pompelor de eflux, printre care: Reserpina ; Fenotiazine: clorpromazina (CPZ), tioridazina (TDZ); Derivative de fenotiazine: 10-(2-Cloropropil)fenotiazina (Brincat et al.); Derivativi de chinazoline (BG1188); Analogi de piperine; Fluorochinolone.

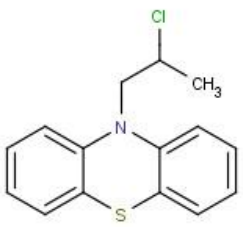
Au fost realizate calcule de structura electronica pentru cateva medicamete studiate in cercetarile experimentale precedente: fenotiazine (clorpromazina, tioridazina, un derivat de fenotiazina (10-(2-cloropropyl)fenotiazina) identificat ca potential substrat intr-un studiu VLS recent, o chinazolina (BG1188) si reserpina, un inhibitor natural al pompelor de eflux in bacterii. A fost calculata geometria optima a compusilor selectati si energia moleculelor optimizate. In Fig. 1 sunt prezentate structurile compusilor chimici studiati iar in Fig. 2 si Fig 3 rezultatele calculelor de optimizare a structurii.



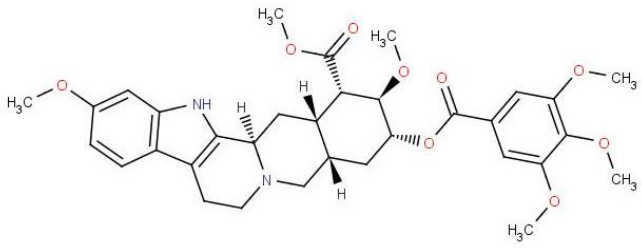
Clorpromazina



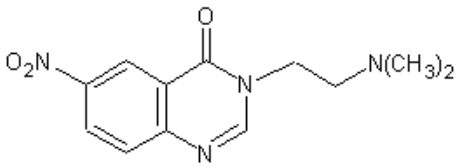
Tioridazina



10-(2-Cloropropil) fenotiazina

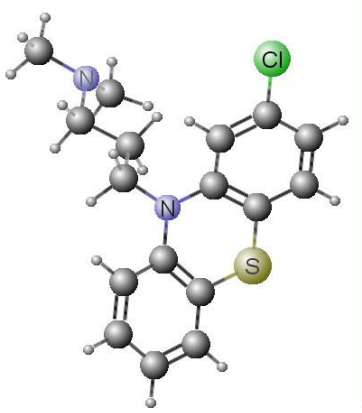


Reserpina

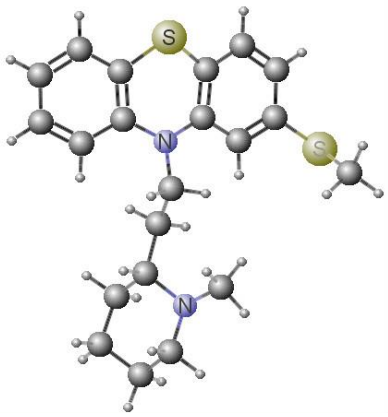


Derivativul chinazolinic BG1188

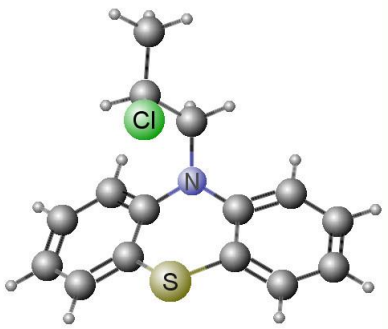
Fig. 1: Structura compusilor EPI propusi pentru studiu



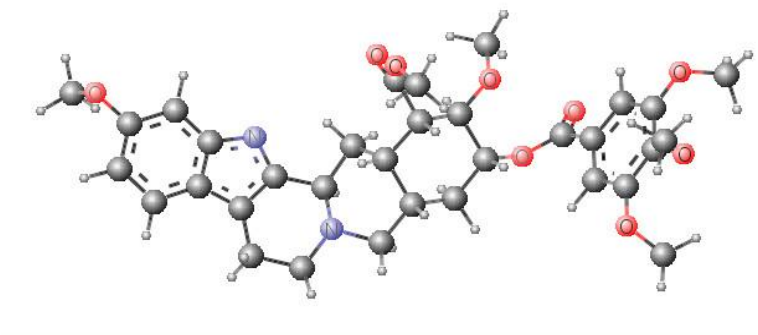
Clorpromazina



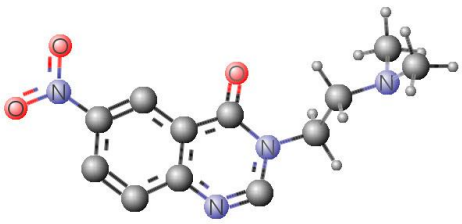
Tioridazina



10-(2-Cloropropil) fenotiazina

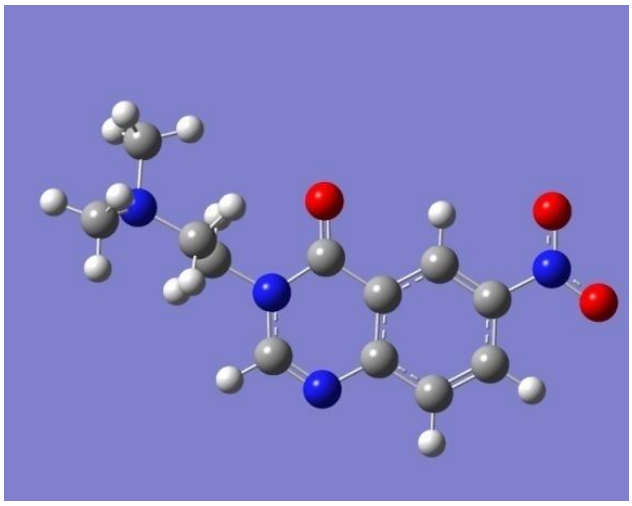


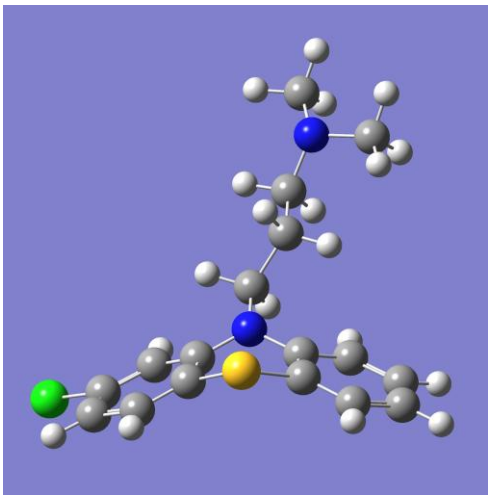
Reserpina



Derivativul chinazolinic BG1188

Fig. 2: Rezultatele optimizarii structurii compusilor EPI studiati





BG1188 (Gaussian09) CPZ (Gaussian 09)

Fig. 3: Rezultatele optimizarii BG1188 si CPZ in Gaussian 09

2. Calcul de structura moleculara

In *S. aureus* au fost identificate mai multe clase de transportori multi drug resistance (MDR):

- ABC: P-glycoprotein, BtuCD, SAV1866, AbcA
- RND: AcrB QacR
- SMR: EmrE, GroEl SepA, QacD
- MATE: NorM MepA
- MFS: EmrD, LacY, GlpT NorA, NorB, NorC, QacA, QacB, MdeA, SdrM

Pentru acest studiu a fost selectata proteina NorA, una din principalele pompe de elfux a *S. Aureus*. A fost prezisa structura proteinei de membrana NorA din *S. Aureus* folosind programul de calcul al impachetarii proteinelor I-Tasser. Alternativ, s-a folosit o proteina de membrana din E. coli, cu un grad de similaritate mare la nivel de secventa de aminoacizi, cu o structura rezolvata din data de difractie cu raze X (LacY). Cele doua structuri sunt prezentate in Fig. 4.

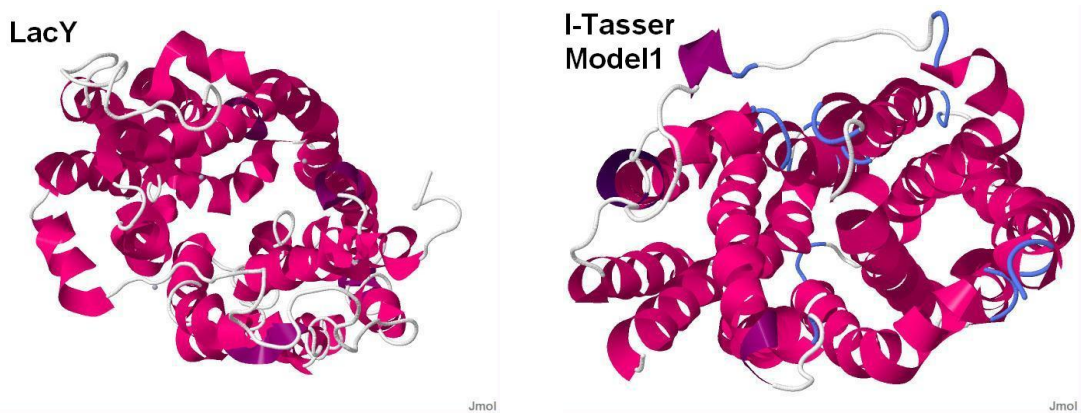


Fig. 4 Modele de impachetare moleculara pentru NorA: LacY si structura prezisa cu I-Tasser.

Docking molecular

Cercetari recente au demonstrat efectul inhibitor al clorpromazinei (CPZ) asupra mai multor pompe bacteriene (Pages et al. 2011). In prezentul studiu am efectuat o analiză comparativă a mai multor compusi chimici si naturali (fenotiazine, chinazoline, reserpina), în scopul de a evalua afinitatea lor pentru norA (proteina rezistenta la chinolone), o pompă de eflux foarte activa in *S. aureus*. Folosind docking molecular cu Autodock4 am estimat afinitatea compusilor chimici selectati pentru un buzunar intern al modelului structural norA (Fig. 5).

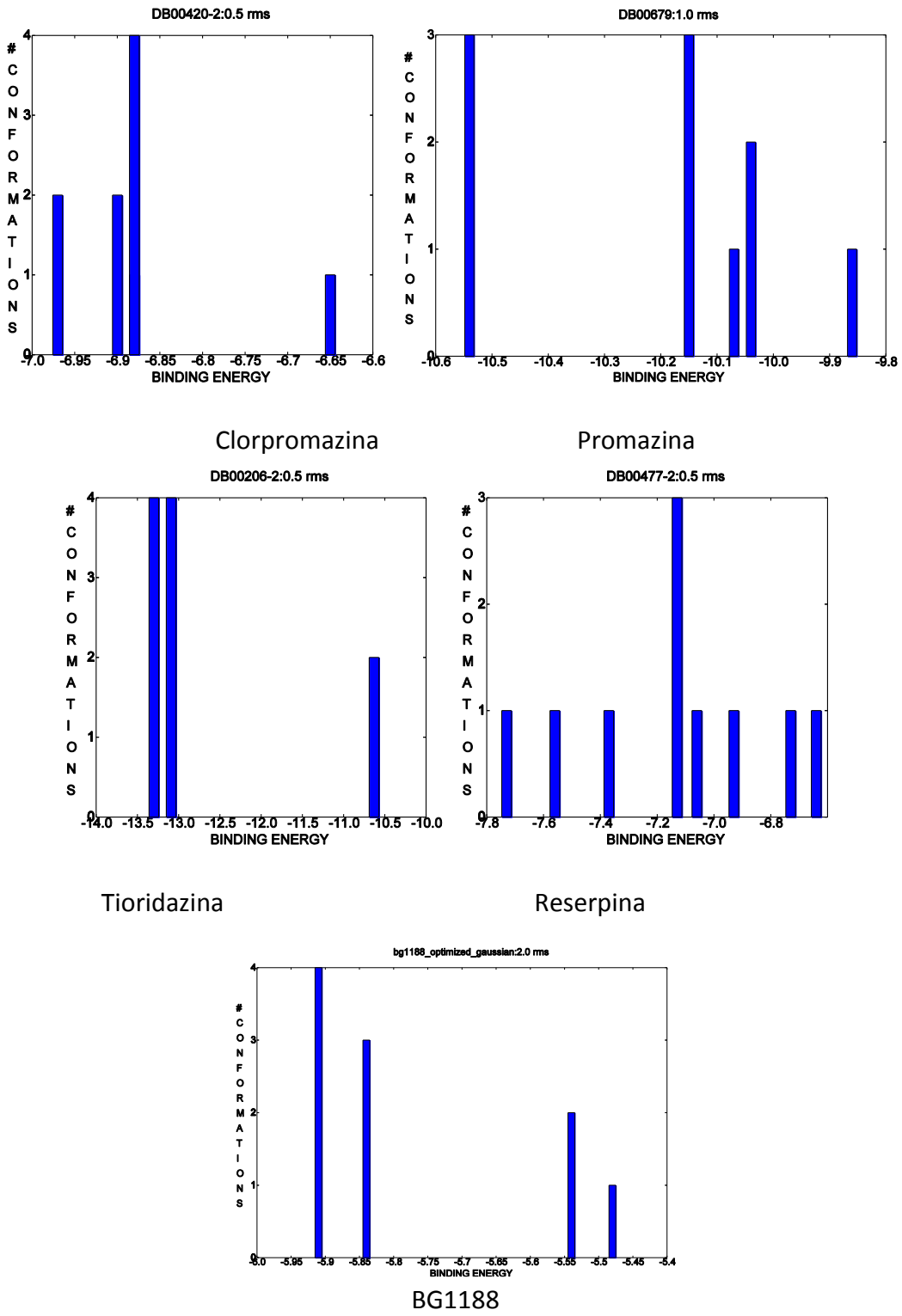


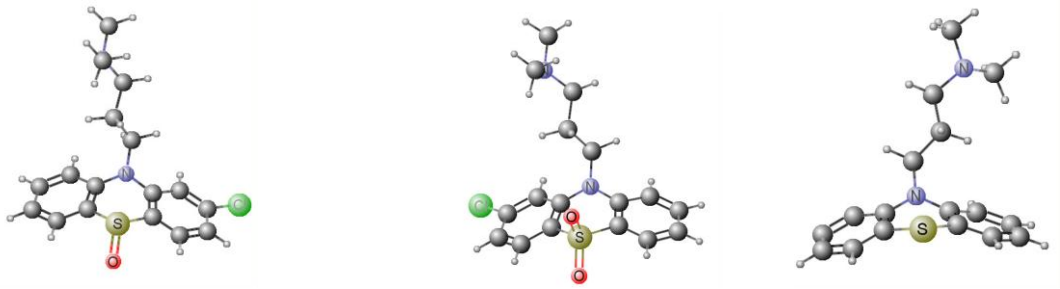
Fig 5: Energia libera Gibbs a sistemului pentru compusi chimici selectati

Modelarea interactiunii radiatiei laser cu CPZ

Caile de reactie posibile in urma iradierii CPZ cu laser in UV includ: fotoionizarea, formarea unui sulfoxid, declorinarea, formarea unui radical promazil.

Au fost studiate, folosind Gaussian, caile de reactie posibile ca urmare a iradierii. Compusii de reactie sunt prezentati in Fig 6. Analiza cailor de reactie va fi comparata cu rezultatele analizei MS a compusilor chimici rezultati in urma iradierii.

Rezultatele detaliate obtinute in cadrul prezentei etape sunt sintetizate in detaliu intr-o propunere de articol data in ANEXA 1.



CPZ SO CPZ SFO CPZ declorinat

Fig. 6 Compusi putativi rezultati in urma iradierii CPZ cu laser in UV

3. Efecte ale expunerii medicamentelor la fascicule laser emise in UV

Spectrul de absorbtie a tioridazinei (TZ) la 10^{-4}M, in apa ultrapura, inainte si dupa iradierea la 355nm si 337nm timp de doua ore este prezentat in fig 7. Se observa ca in cazul expunerii la 355nm, atat maximul de absorbtie la 262nm cat si intensitatea acestuia nu sunt modificate. In contrast, expunerea la 337,1 nm produce o deplasare cu 3nm si o reducere cu 30% a intensitatii maximului de absorbtie. De asemenea expunerea la 337,1nm timp de doua ore urmata de iradierea timp de 30 de minute la 355nm produce un efect asemanator iradierii cu 337,1nm. Iradierea timp de 4 ore la 266nm produce o deplasare a maximului de absorbtie cu 6nm fata de proba neiradiata si o scadere cu 40% in intensitate a acestuia.

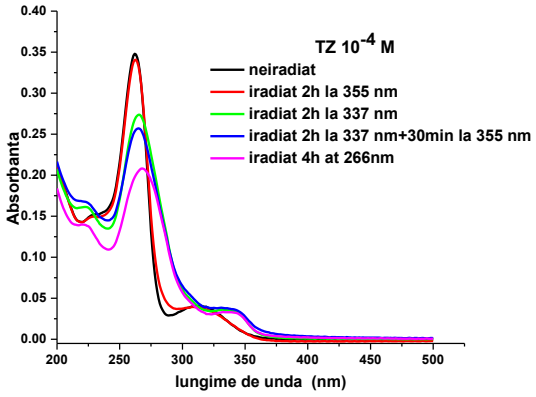


Fig. 7**.** Spectrele de absorbtie ale solutiilor de tioridazina 10^{-4}M in apa ultrapura inainte si dupa expunerea la radiatie laser de diferile lungimi de unda.

Expunerea tioridazinei in concentratie de $5x10^{-2}$M timp de o ora la 337,1nm produce un al doilea maxim de absorbtie la aproximativ 650nm, acesta estompandu-se dupa depozitatea probei iradiate timp de 19 ore la o temperatura de 4°C. Expunerea simultana la lungimile de unda 266 si 532nm nu a produs modificarile observate in cazul anterior.

Spectre de fluorescenta indusa laser au fost inregistrate in timp real prin masurarea probelor neiradiate sau a celor iradiate in prealabil cu 337,1nm, 266nm, 532nM si 266nm&532nm aplicate simultan fig 8.

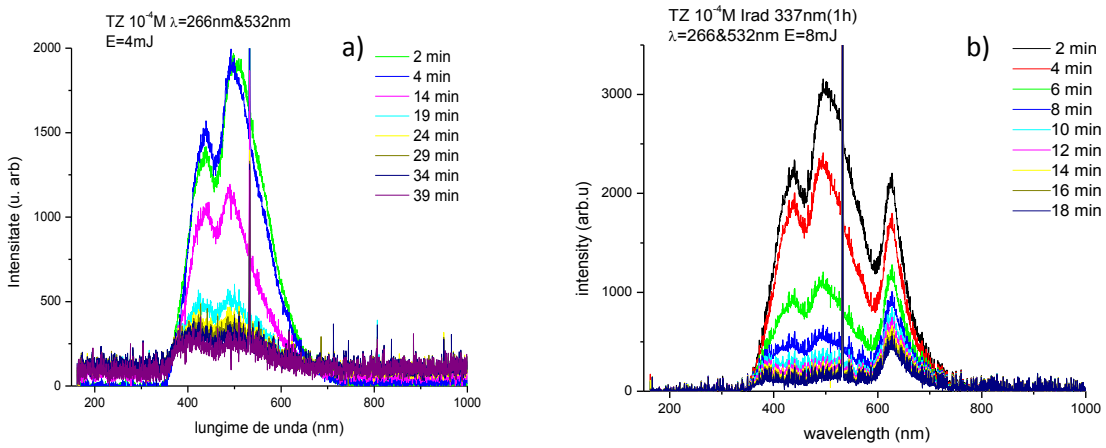


Fig. 8. Fluorescenta indusa laser a tioridazinei a) neiradiata si b) iradiata in prealabil cu 337,1nm

Tulpina de referinta *Staphylococcus aureus* este rezistenta la Clorpromazina (CPZ) dupa cum se poate observa din absenta unei zone de inhibitie a cresterii bacteriei chiar si in cazul in care au fost aplicate 500µg de CPZ neiradiat, dupa cum se poate observa in fig 9.1.

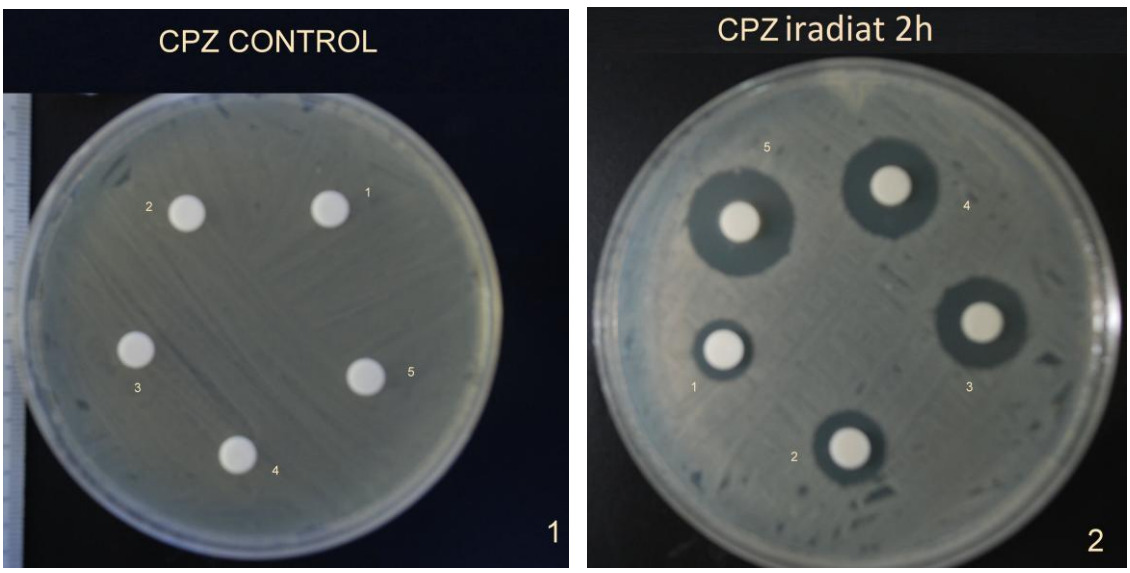


Fig. 9. Proba biologica pentru determinarea activitatii CPZ asupra cresterii tulpinii de *Staphylococcus aureus.*

Deoarece expunerea clorpromazinei la alte lungimi de unda nu afecteza cresterea bacteriilor, s-au efectuat iradieri timp de doua ore la 266nm, in urma carora s-au obtinut produsi de reactie care inhiba cresterea bacteriei (Figura 9.2).

Bibliografie:

Autodock 4, online: http://autodock.scripps.edu/

Militaru, A., A. Smarandache, A. Mahamoud, V. Damian, P. Ganea, S. Alibert, J. M. Pages, and M. L. Pascu. Stability Characterization of Quinazoline Derivative BG1188 by Optical Methods, AIP Conf. Proc. 1364, 13 (2011).

Militaru, A., Smarandache, A, Mahamoud, A., Alibert, S., Pages, J. M., Pascu, M. L. Time Stability Studies of Quinazoline Derivative Designed to Fight Drug Resistance Acquired by Bacteria Letters in Drug Design & Discovery, Volume 8, Number 2, February 2011 , pp. 124-129(6).

Pascu, M. L., Nastasa V., Smarandache A., Militaru A., Martins A., Viveiros M., Boni M., Andrei I. R., Pascu A., Staicu A., Molnar J., Amaral L., (2011) Direct modification of bioactive phenothiazines by exposures to laser radiation, to be published in *Recent Patents on Anti-Infective Drug Discovery*.

Popescu, G.V., Militaru A., Pascu M. L., Nastasa V., Staicu A., "Comparative docking of several phenothiazines and quinazolines with resistant strains of as inhibitors of Staphylococcus aureus NorA ," submission in preparation.

Schmidt M. *et al.*, (1993) General atomic and molecular electronic structure system, J. Comput. Chem. 14, 1347

RAPORT STIINTIFIC SI TEHNIC

DFCTI/IFIN-HH

REZUMATUL ETAPEI

I. RAPORT TEHNIC

In prima etapa activitatea grupului din DFCTI/IFIN-HH s-a concentrat asupra dezvoltarii unei platforme de calcul de inalta performanta (High Performance Computing - HPC) necesara pentru modelarea si analiza compusilor macromoleculari ce vor fi investigati in cadrul proiectului. Platforma urmeaza sa asigure suportul hardware si software necesar pentru determinarea structurii proteinelor, pentru cautarea secventelor genomice asemanatoare si a domeniilor similare in proteine, pentru modelare moleculara, precum si pentru simularea dinamicii moleculare.

Arhitectura de ansamblu a sistemului de modelare si simulare, care include platforma realizata la IFIN-HH, este reprezentata in Fig. 1.

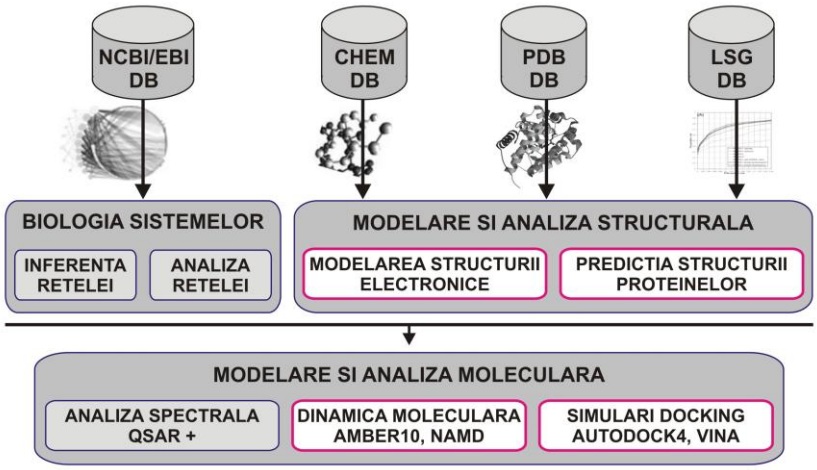


Fig.1: Arhitectura sistemului de modelare si simulare AMPLE

Sistemul utilizeaza informatii din baze de date (DB) proprii (cum este cea a grupului de spectroscopie laser – LSG din INFLPR) sau publice, inclusiv cele oferite de NCBI (National Center for Biotechnology Information [1]), EBI (European Biology Institute [2]), Univ. of California – Irvine (ChemDB [3]), sau Protein Data Bank (PDB [4]).

In IFIN-HH, platforma HPC - ale carei functii principale sunt reprezentate pe fond deschis in Fig.1 - a fost implementata prin actualizarea unor componente ale bazei de calcul paralel existenta in DFCTI si adaugarea de elemente hardware si software noi, adaptate cerintelor specifice ale proiectului. Aceasta include un cluster de calcul paralel pentru calcule intensive, o statie de lucru pentru predictii de structura, analiza si reprezentarea grafica a datelor, si un server care gazduieste bazele de date.

Din punct de vedere software, s-a avut in vedere realizarea unei structuri deschise si flexibile, bazata atat pe solutii open source cat si proprietare, care sa permita adaugarea ulterioara a unor noi componente software in masura in care acestea se vor dovedi necesare.

In stadiul actual, platforma contine urmatoarele instrumente software:

- Gaussian 9.0 (http://www.gaussian.com/) – pentru calcule de structura electronica [5];

- I-TASSER 2.1 (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) - pentru predictia structurii proteinelor [6];

- BLAST [7] si mpiBLAST 1.6.0 (http://www.mpiblast.org/) – pentru cautari de similaritate in secvente;

- Amber 12.0 [8], CHARMM [9] si NAMD [10] – pentru modelare moleculara si simulari de dinamica moleculara.

Activitatile desfasurate si rezultatele obtinute sunt detaliate mai jos.

II. RAPORT STIINTIFIC

In cadrul primei etape a proiectului s-a elaborat si s-a trimis spre publicare la J. of Chemical Information and Modeling lucrarea *'Building a Knowledge-based Statistical Potential by Capturing High-Order Inter-Residue Interactions and its Applications in Protein Secondary Structure Assessment'*, autori Yaohang Li, Hui Liu, Ionel Rata, si Eric Jakobsson, care se afla in acest moment in ultima faza de review.

In continuare se va face o scurta prezentare a acestei lucrari si a relevantei acesteia pentru proiectul AMPLE.

Una dintre sarcinile importante ale proiectului consta in modelarea structurii proteinelor transmembranare ce produc e-fluxul antibioticelor in bacteriile rezistente. Dupa aflarea genei corespunzatoare aceasta poate fi transpusa in secventa primara a aminoacizilor constituenti. Apoi, din structura primara trebuie aflata structura tertiara cu un grad cat mai mare de precizie pentru a face posibila simularea ulterioara a inhibarii proteinei prin dockingul cu compusi organici. Aceasta este o sarcina foarte dificila si nici cele mai performante programe nu reusesc sa atinga un grad de precizie mai mare decat aproximativ 5A rmsd. Metoda cea mai utilizata in acest caz este modelarea prin homologie cu o proteina similara, a carei structuri este cunoscuta experimental. Modelarea prin homologie de obicei detrmina pozitiile structurilor helicoidale transmembranare ale proteinei necunoscute prin suprapunere cu cea cunoscuta.

Lucrarea de fata are scopul de a face o verificare mai atenta a preciziei de determinare a structurilor helicoidale transmembranare pentru a sti exact unde incep si unde se sfarsesc acestea, si a determina implicit lungimile si secventele exacte ale buclelor exterioare dintre ele ce formeaza gura de intrare a canalului proteinei. Aceasta gura de intrare este portiunea unde se realizeaza de obicei docking-ul.

Lucrea este prima dintr-o serie de incercari de a crea potentiale ce analizeaza statistic informatii culese din baze de date structurale (PDB), pentru a furniza noi masuri ale interactiilor moleculare de interes. Aceasta lucrare ofera o analiza statistica a structurilor secundare din PDB, pe baza careia se pot face predictii si analize de structuri secundare necunoscute. Metoda este inedita in domeniul determinarii de structuri secundare si s-a dovedit a fi foarte precisa in testarile efectuate. In 56% dintre proteinele testate, metoda prezentata in lucrare este capabila sa prezica structurile secundare native cu o precizie mai mare de 90%, iar in mai mult de 80% din teste, precizia este de peste 80%. Astfel, de la prima incercarea autorilor in acest domeniu foarte popular, acestia au reusit sa obtina rezultate comparabile cu cele mai bune metode existente. Metoda este inca perfectibila si este de asteptat ca aceste rezultate sa se imbunatateasca in continuare.

Metoda functioneaza cu precizie in special in cazul structurilor helicoidale si in deosebi in cele transmembranare pentru care delimitarile capetelor sunt mai bine semnalate de suprafetele membranei. Pentru proiectul AMPLE ne propunem sa efectuam analiza statistica doar pe un set de proteine membranare din PDB si sa consideram doar structurile helicoidale si buclele ca elemente de structuri secundare. Numarul acestor elemente se reduce astfel la 2, ceea ce va imbunatati analiza statistica.

Potentiale statistice similare pot fi create si pentru alte analize si predictii de structura necesare pentru proiect. Avantajul potentialelor statistice este ca sunt mai flexibile decat potentialele fizice si pot folosi orice proprietati considerate semnificative problemei specifice. In final, este de semnalat faptul ca prin aceasta lucrare autorii au fost primii care au aplicat cu succes metoda potentialelor statistice la problema prezicerii structurii secundare a proteinelor.

Lucrarea este reprodusa integral in ANEXA la acest raport.

Building a Knowledge-based Statistical Potential by Capturing High-Order Inter-Residue Interactions and its Applications in Protein Secondary Structure Assessment

Yaohang Li[*,1], *Hui Liu*[2], *Ionel Rata*[3], *and Eric Jakobsson*[4]

[1]Department of Computer Science

Old Dominion University

yaohang@cs.odu.edu

[2]Center for Biophysics and Computational Biology

University of Illinois at Urbana-Champaign

huiliu2@illinois.edu

[3]National Institute for Physics and Nuclear Engineering (IFIN-HH), R-77125, Bucharest-Magurele, Romania

ionel.rata@nipne.ro

[4]Department of Molecular and Integrative Physiology, Beckman Institute, and National Center for Supercomputing Applications

University of Illinois at Urbana-Champaign

jake@ncsa.illinois.edu

ABSTRACT. The rapidly increasing number of protein crystal structures available in PDB has naturally made statistical analyses feasible in studying complex high-order inter-residue correlations. In this paper, we report a Context-based Secondary Structure Potential (CSSP) for assessing the quality of predicted protein secondary structures generated by various prediction servers. CSSP is a sequence-position-specific knowledge-based potential generated based on the potentials of mean force approach, where high-order inter-residue interactions are taken into consideration. The CSSP potential is effective in identifying secondary structure predictions with good quality. In 56% of the targets in the CB513 benchmark, the optimal CSSP potential is able to recognize the native secondary structure or a prediction with Q3 accuracy higher than 90% as best scored in the predicted secondary structures generated by 10 popularly used secondary structure prediction servers. In more than 80% of the CB513 targets, the predicted secondary structures with the lowest CSSP potential values yield higher than 80% Q3 accuracy. Moreover, our computational results also show that the CSSP potential using triplets outperforms the CSSP potential using doublets and is currently better than the CSSP potential using quartets.

1. Introduction

Prediction of protein secondary structure from the primary sequence is an important step toward prediction of tertiary structure. The more accurately the secondary structure can be predicted, the smaller the search space for the tertiary structure prediction. At the core of the secondary structure prediction problem is the derivation of knowledge for secondary structure assignment. The knowledge is contained in the Protein Data Bank (PDB), which includes 83,983 protein structures as of Aug. 21, 2012, specifically in the secondary structure assignment as reported in the PDB. Nevertheless, generation of knowledge for secondary structure assignment is complicated by several sources of inherent error. In the first place, the tertiary structure from which the secondary structure is derived has a resolution ranging from one to a few angstroms, sufficient to alter the local secondary structure assignment. Secondly, the algorithms that translate the tertiary structure to a secondary structure necessarily have a tolerance for a range of backbone torsion angles that define any of the well-defined secondary structures. These two bases for uncertainty about the precise secondary structure of proteins in PDB contribute to the fact that the maximum meaningful secondary structure prediction accuracy that can ever be obtained, given the noise in the experimental data and its analysis, is significantly less than 100%. It has been estimated at about 88%~90%[24].

Pirovano and Heringa[25] have recently done a critical comparative study of protein secondary structure prediction methods. By the metrics they use in their study, which are generally consistent with other studies and with our group's experience (unpublished), the existing methods provide accuracies near 80%. Wei et al.[26] have utilized linear optimization to provide weighting for a consensus prediction of seven different methods. They report consensus predictions have averagely a couple of percent better than the best single method, suggesting that a consensus method may move the state of the art a significant fraction towards the theoretical maximum, but still far short of

the theoretical maximum. As a basis for tertiary structure prediction, moving the percent of inaccuracy from the high teens to 10 percent would be an enormous improvement in efficiency, because the search space for finding a tertiary structure goes up superlinearly with the fraction of inaccuracy in the secondary structure prediction. Because of a combinatorial expansion of possibilities, such an improvement in secondary structure prediction would reduce the search space for predicting tertiary structure many-fold.

In the present paper, we describe an approach of integrating knowledge for secondary structure assignment into a knowledge-based potential to assess the quality of predicted secondary structures. We hypothesize that incorporating higher-order inter-residue correlations into the knowledge-based potential is likely to lead to high accuracy. In particular, we note that it is reasonable to expect correlations of identity for pairs of residues one position removed from each other in turns, two positions removed from each other in β-strands, and three and four positions removed from each other in helices.

When there were relatively few experimental structures available, capturing high-order inter-residue interactions into knowledge-based potentials was difficult due to lack of statistical samples. We note that the sample size for specific doublets in the PDB is 1/20 of that for individual residues (singlets), for specific triplets is 1/20 of that for doublets, and for quartets 1/20 of that for triplets. The fractions are even smaller if rare amino acids are involved. However recently, as an increasing number of high-resolution protein crystal structures are available in Protein Data Bank (PDB), and powerful computers are available to sort through larger dimension combinatory, it has become feasible to derive knowledge for high-order inter-residue interactions and incorporate it into a knowledge-based potential.

Most of the secondary structure prediction methods[8-17] consider inter-residue correlation implicitly by encoding a window of 15-21 residues in neural networks or other learning machines. Although these methods have achieved certain success, the neural networks or learning machines

work like "black boxes," which provide little understandable information in the relation between inter-residue interactions and the secondary structures. Only a few methods have attempted to estimate (high-order) inter-residue correlations explicitly. Miyazawa and Jernigan[29] developed a secondary structure energy using the potentials of mean force method by considering the three-body interactions among three consecutive residues. The GOR4[7] method treats inter-residue interactions as information functions of events and integrates them according to the information theory. The original GOR4 program only considers singlets and doublets within a window. The later GOR5 program[27] takes higher order interactions such as triplets into account and but finds that the improvement is only 0.3%. The authors suggest that "better optimization and larger database" are necessary for further accuracy improvement[27]. More recently, Madera et al.[28] proposed a simple *k*-mer model using a conditional random field to achieve more "realistic" secondary structure predictions.

In this paper, we derive statistics of singlets, doublets, triplets, and quartets of residues with specific relative occurrences at sequence positions and then convert them to inter-residue interaction potentials using the potentials of mean force[1] approach. A Context-based Secondary Structure Potential (CSSP) integrating these inter-residue interaction potentials is developed for assessing predicted protein secondary structures. We use the cull datasets (CullPDB) generated by the PISCES server[3] as the training sets for CSSP. We test CSSP by using it to evaluate the predicted secondary structures generated by 10 public secondary structure prediction servers, including GOR4[7], HNN[8], SAM[9], Jpred[10], Psipred[12], ProfPHD[11, 13], Jufo[14], Netsurfp[15], SSPRO4[16], and Porter[17], using a commonly used set of sequences known as the CB513 benchmark[5]. For the correctness of our computational experiments, chains in CB513 and their homologs are removed from the CullPDB to ensure the separation of training set and testing set. Accuracy comparisons of potentials with different orders and ranges of inter-residue interactions are also made.

2. Methods

2.1 Knowledge-based Statistical Potential for N-residue Fragments with High-Order Inter-Residue Interactions

2.1.1 Formation of the *k*-let Potential

Our formation of the potential is based on the mean-force potential energy according to the Boltzmann formula[1]. We firstly come up with a statistical potential for a *k*-let at residue positions $i_1, i_2, \dots, i_k$ in a protein sequence. The derivation of a statistical potential $U\left(R_{i_1}R_{i_2}\dots R_{i_k}, C_{i_1}C_{i_2}\dots C_{i_k}\right)$ for a sequence-structure correlated *k*-let starts from the common form of statistical potential calculation using inverse Boltzmann theorem:

$$U\left(R_{i_1}R_{i_2}\dots R_{i_k}, C_{i_1}C_{i_2}\dots C_{i_k}\right) = -RT\ln\frac{P_{obs}(C_{i_1}C_{i_2}\dots C_{i_k}|R_{i_1}R_{i_2}\dots R_{i_k})}{P_{ref}(C_{i_1}C_{i_2}\dots C_{i_k}|R_{i_1}R_{i_2}\dots R_{i_k})}, \quad i_1 \neq i_2 \neq \dots \neq i_k$$

where $P_{obs}(C_{i_1}C_{i_2}\dots C_{i_k}|R_{i_1}R_{i_2}\dots R_{i_k})$ is the observed probability of *k*-let $R_{i_1}R_{i_2}\dots R_{i_k}$ with conformation $C_{i_1}C_{i_2}\dots C_{i_k}$, $P_{ref}(C_{i_1}C_{i_2}\dots C_{i_k}|R_{i_1}R_{i_2}\dots R_{i_k})$ is the probability of the reference state, *R* is the gas constant, and *T* is the temperature. Using the frequency values to estimate the probability $P_{obs}(C_{i_1}C_{i_2}\dots C_{i_k}|R_{i_1}R_{i_2}\dots R_{i_k})$ and applying the conditional probability method described in Samudrala and Moult[2], $U\left(R_{i_1}R_{i_2}\dots R_{i_k}, C_{i_1}C_{i_2}\dots C_{i_k}\right)$ can be written as

$$U\left(R_{i_1}R_{i_2}\dots R_{i_k}, C_{i_1}C_{i_2}\dots C_{i_k}\right) = -RT\ln\left.\frac{N_{obs}(C_{i_1}C_{i_2}\dots C_{i_k}, R_{i_1}R_{i_2}\dots R_{i_k})}{N_{obs}(R_{i_1}R_{i_2}\dots R_{i_k})}\right/\frac{N_{obs}(C_{i_1}C_{i_2}\dots C_{i_k})}{N_{total}}$$

where $N_{obs}(C_{i_1}C_{i_2}\dots C_{i_k}, R_{i_1}R_{i_2}\dots R_{i_k})$ is the observed number of *k*-let $R_{i_1}R_{i_2}\dots R_{i_k}$ with conformation $C_{i_1}C_{i_2}\dots C_{i_k}$ in a protein structure database, $N_{obs}(R_{i_1}R_{i_2}\dots R_{i_k})$ is the number of observations of $R_{i_1}R_{i_2}\dots R_{i_k}$, $N_{obs}(C_{i_1}C_{i_2}\dots C_{i_k})$ is the number of observations of $C_{i_1}C_{i_2}\dots C_{i_k}$, and N_{total} is the total number of observations.

Two *k*-lets are of the same kind if their residues positions $i_1, i_2, \ldots, i_k$ and $i'_1, i'_2, \ldots, i'_k$ (in the same or different protein sequences) have the same relative sequence distances: $i_1 - i'_1 = i_2 - i'_2 = \cdots = i_k - i'_k$. Then, $N_{obs}(C_{i_1}C_{i_2}\ldots C_{i_k}, R_{i_1}R_{i_2}\ldots R_{i_k})$ can be obtained by counting the total number of occurrences of *k*-lets having conformation $C_{i_1}C_{i_2}\ldots C_{i_k}$ at the same relative residue positions as $i_1, i_2, \ldots, i_k$ in the protein structure database. Similar calculations can be applied to obtain $N_{obs}(R_{i_1}R_{i_2}\ldots R_{i_k})$, $N_{obs}(C_{i_1}C_{i_2}\ldots C_{i_k})$, and N_{total}.

For simplicity, we use $U(i_1, i_2, \ldots, i_k)$ to represent the *k*-let potential $U\left(R_{i_1}R_{i_2}\ldots R_{i_k}, C_{i_1}C_{i_2}\ldots C_{i_k}\right)$ in the rest of the paper.

2.1.2 Interaction Potential

We denote $INT(i_1, i_2)$ to capture the two-body (doublet) interaction potential energy between residues R_{i_1} and R_{i_2}

$$INT(i_1, i_2) = U(i_1, i_2) - U(i_1) - U(i_2).$$

Similarly, the higher order three-body interactions $INT(i_1, i_2, i_3)$ of triplet residues R_{i_1}, R_{i_2}, and R_{i_3} can be expressed as

$$INT(i_1, i_2, i_3) = U(i_1, i_2, i_3) - U(i_1) - U(i_2) - U(i_3) - INT(i_1, i_2) - INT(i_1, i_3) - INT(i_2, i_3).$$

For a *k*-let, the high order *k*-body interactions $INT(i_1, i_2, \ldots, i_k)$ of residues $R_{i_1}R_{i_2}\ldots R_{i_k}$ can be generalized as

$$\begin{aligned} INT(i_1, i_2, \ldots, i_k) &= U(i_1, i_2, \ldots, i_k) - \sum_{j=1}^{k} U(i_j) - \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k} INT(i_{j_1}, i_{j_2}) \\ &- \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k}\sum_{j_3=j_2+1}^{k} INT(i_{j_1}, i_{j_2}, i_{j_3}) - \cdots \\ &- \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k} \cdots \sum_{j_{k-1}=j_{k-2}+1}^{k} INT(i_{j_1}, \ldots, i_{j_{k-1}}). \end{aligned}$$

2.1.3 Potential of *N*-residue Fragment

By considering up to *k*-body interactions, we can represent the mean-force potential $U(M+1, M+2, \ldots, M+N)$ of an *N*-residue fragment $R_{M+1}R_{M+2}\ldots R_{M+N}$ starting at the (*M+1)*th position in a protein sequence as

$$\begin{aligned} &U(M+1, M+2, \ldots, M+N) \\ &\quad = \sum_{j=1}^{N} U(M+j) + \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} INT(M+j_1, M+j_2) \\ &\quad + \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\sum_{j_3=j_2+1}^{N} INT(M+j_1, M+j_2, M+j_3) + \cdots \\ &\quad + \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\cdots\sum_{j_k=j_{k-1}+1}^{N} INT(M+j_1, M+j_2, \ldots, M+j_k). \end{aligned}$$

By substituting the interaction potential with *k*-let potential and combining the common terms, the potential energy of an *N*-residue fragment $R_{M+1}R_{M+2}\ldots R_{M+N}$ is simplified as the weighted sum of potentials of singlets, doublets, triplets, …, and up to *k*-lets.

$$\begin{aligned} &U(M+1, M+2, \ldots, M+N) \\ &\quad = \underbrace{w_1 \sum_{j=1}^{N} U(j)}_{singlet} + \underbrace{w_2 \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} U(M+j_1, M+j_2)}_{doublet} \\ &\quad + \underbrace{w_3 \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\sum_{j_3=j_2+1}^{N} U(M+j_1, M+j_2, M+j_3) + \cdots}_{triplet} \\ &\quad + \underbrace{w_s \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\cdots\sum_{j_s=j_{s-1}+1}^{N} U(M+j_1, M+j_2, \ldots, M+j_s) + \cdots}_{s-let} \\ &\quad + \underbrace{w_k \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\cdots\sum_{j_k=j_{k-1}+1}^{N} U(M+j_1, M+j_2, \ldots, M+j_k)}_{k-let} \end{aligned}$$

where the weights w_s are

$$w_s = \sum_{j=s}^{k} (-1)^{j-s} \binom{N-k}{j-s}.$$

Using the potential of *N*-residue fragments and removing the overlapping parts, the overall potential energy of a protein with *L* residues is

$$\begin{aligned} U_{protein} &= U(1,\dots,N) + U(2,\dots,N+1) - U(2,\dots,N) + U(3,\dots,N+2) - U(3,\dots,N+1) + \cdots \\ &= \sum_{j=1}^{L-N+1} U(j,\dots,j+N-1) - \sum_{j=2}^{L-N+1} U(j,\dots,j+N-2) \end{aligned}$$

2.2 Potentials for Secondary Structure Prediction Evaluation

2.2.1 Datasets

We use the CullPDB datasets generated by the PISCES server[3] to collect *k*-let samples to produce CSSP potentials to evaluate secondary structure predictions. The CullPDB datasets generated on 10/21/2011 with maximum 3.0A resolution and maximum 1.0 R-factor are selected. A public benchmark CB513 is used as a testing set to validate our methods. To ensure the correctness of our computational experiments, we enforce the separation of training set and testing set by excluding all sequences with greater than 25% identity to any sequence in CB513 from the CullPDB datasets when the *k*-let samples are extracted to calculate the statistical potential. Moreover, the *k*-let samples with missing residues are discarded. Furthermore, due to the fact that PSI-BLAST is usually unable to generate profiles for short sequences, the protein sequences with lengths less than 30 are also removed from the CullPDB datasets.

2.2.2 Estimation of *k*-let Probability

The weighted frequency value of a *k*-let of a certain secondary structure appeared in the CullPDB is used to estimate the probability of the *k*-let sample adopting this secondary structure. The weights of *k*-let samples are based on the PSSM (Position Specific Score Matrix) frequencies at each residue position. PSSM data contains evolutionary information derived from sequence

homologues. For a given protein in the CullPDB datasets, PSI-BLAST[4] is used to search against the NR (Non-Redundant) database with E-value = 0.001 and at most 3 iterations. After the PSSM file is generated, weights are calculated according to the frequency of each residue appearing in a specific position of the sequence. For example, the following figure shows a segment of a PSSM frequency table where a four-residue fragment "ASYK" has secondary structure of "HHHC".

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
												...									
1	A	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0
2	S	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	47	0	0	0
3	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	49	0	0	0	0	51	0
4	K	1	9	6	2	0	13	10	0	2	0	0	34	0	0	1	6	16	0	0	0
												...									

Then, in triplet calculation, weight of 88/100*24/100*49/100 = 0.103 is counted toward $N_{obs}(HHH,AAF)$ at triplet position "1_2_3", weight of 88/100*47/100*9/100 = 0.037 is counted toward $N_{obs}(HHH,ATR)$ at triplet position "1_2_4", weight of 12/100*51/100*34/100 = 0.021 is counted toward $N_{obs}(HHC,SYK)$ at triplet position "1_3_4", and so on. These combinations give many samples with different weights for calculating the frequency of *k*-lets at different positions.

3. Results

3.1 CSSP using Triplets

We firstly investigate the sensitivity and accuracy of our new knowledge-based statistical potential CSSP by incorporating three-body inter-residue (triplet) interactions using the CB513 benchmark. Since CSSP does not include statistics for unidentified residues, we only consider the 507 out of the 513 targets in CB513 benchmark excluding the 6 with unidentified residues in the protein sequence. We create a secondary structure set composed of the predicted secondary structures from 10 public prediction servers as well as the native structure and test if the knowledge-based potential can recognize the high quality predictions. The 10 prediction servers we used include GOR4, HNN, SSPRO4, PORTER, NETSURFP, PSIPRED, SAM, PROFPHD, and JUFO.

The precisions of the predicted secondary structures are measured by the Q3 accuracy, i.e., the total accuracy of three classes – α-helix, β-strand, and coil. These predicted secondary structure conformations have very different qualities. GOR4 is an early statistical model based on frequencies of amino acid pairs and HNN is an early method using neural network for classification – both have relatively low accuracy compared to the modern secondary structure prediction servers. On the other hand, SSPRO4 and PORTER take advantage of the homologue structural information for prediction. When structures of homologues with 50% or higher sequence identity are available, SSPRO4 or PORTER can often produce high quality predictions with Q3 accuracy around 80%~90% or even 100%[6]. NETSURFP, PSIPRED, SAM, PROFPHD, JUFO, and JPRED are popularly used prediction servers using neural networks[8, 13, 16, 17], hidden Markov Chains[9], Support Vector Machine (SVM)[12], or consensus[14] methods, typically having Q3 accuracy between 70% and 80%. Table 1 compares the performance of the 10 public servers for secondary structure prediction on CB513.

Methods	# of Targets with Q3 > 90%	# of Targets with Q3 > 80%
GOR4	1 (0.20%)	20 (3.94%)
HNN	8 (1.58%)	40 (7.89%)
NETSURFP	29 (5.72%)	230 (45.36%)
PSIPRED	58 (11.44%)	327 (64.50%)
SAM	25 (4.93%)	226 (44.58%)
SSPRO4	432 (85.21%)	468 (92.31%)
PORTER	453 (89.35%)	492 (97.04%)
PROFPHD	10 (1.97%)	129 (25.44%)
JPRED	32 (6.31%)	269 (53.06%)
JUFO	5 (0.99%)	126 (24.85%)

Table 1. Performance of 10 Secondary Structure Prediction Methods on CB513

In this paper, we measure the identification accuracy of the CSSP potentials by the percentage of targets in CB513 in which the predicted structures yielding the lowest potential energy values have Q3 accuracies higher than 80% or 90%. Because the secondary structure assignments based on crystal structure have ~10% errors themselves[20, 21] as inferred from

differences between different X-ray structures and NMR models of the same protein and from inconsistency of secondary structure assignments by different methods of different parameters, e.g., DSSP[22] and STRIDE[23], 90% Q3 prediction accuracy is usually considered as the upper bound of secondary structure prediction. Predictions with 80% Q3 accuracies are also regarded as models with high precision.

A number of tests have been carried out to determine the optimal parameters for CSSP using triplets, including the Cull datasets, the fragment size, and the number of iterations in PSI-BLAST.

Figure 1 compares the identification accuracies when Cull datasets with maximum pairwise mutual sequence identity ranging from 20% to 90% are used to generate the CSSP potentials with fragment size 7 in CB513. On one hand, datasets with lower sequence identity have fewer protein sequences and thus fewer triplet samples. On the other hand, samples may bias to certain protein families in datasets with higher sequence identity. Figure 1 shows that the Cull dataset with maximum 50% sequence identity have the best compromise of sampling accuracy by showing the highest overall identification percentages. For the Cull dataset with maximum 50% sequence identity, in 56.2% and 80.1% of the CB513 targets, CSSP can pick up one from the 10 predicted structures generated by the prediction servers having higher than 90% and 80% Q3 accuracy, respectively.

Figure 2 shows the overall Q3 accuracy in CB513 of varying fragment sizes using CSSP trained by the Cull dataset with maximum 50% sequence identity. The CSSP with fragment size of 7 yields the best result, with overall Q3 accuracy of 88.2%. The optimum fragment size of 7 has certain biological meaning – triplet residues in helix, strand, and coil are strongly correlated at relative positions 1-3-5, 1-4-7, and 1-2-3, respectively. For bigger fragment sizes than 7, the identification accuracies drop gradually, due to the reason that the importance of long distance inter-residue correlation decreases while the statistical sampling noise accumulates.

Since CSSP takes advantage of the evolutionary information to generate the statistics for *k*-lets, the evolutionary distance occupied by a protein and its homologs also affects the accuracy of CSSP. Figure 3 investigates the accuracy of CSSP using weighted frequencies generated from PSSM using 3 and 6 PSI-BLAST iterations. One can find that both CSSPs yield similar performance, but the one based on PSI-BLAST using 3 iterations is slightly more sensitive. This may be due to the fact that more PSI-BLAST iterations bring in more less-related homologs in the protein family with likely more diverse structures, which reduces the sensitivity of the *k*-let statistics.

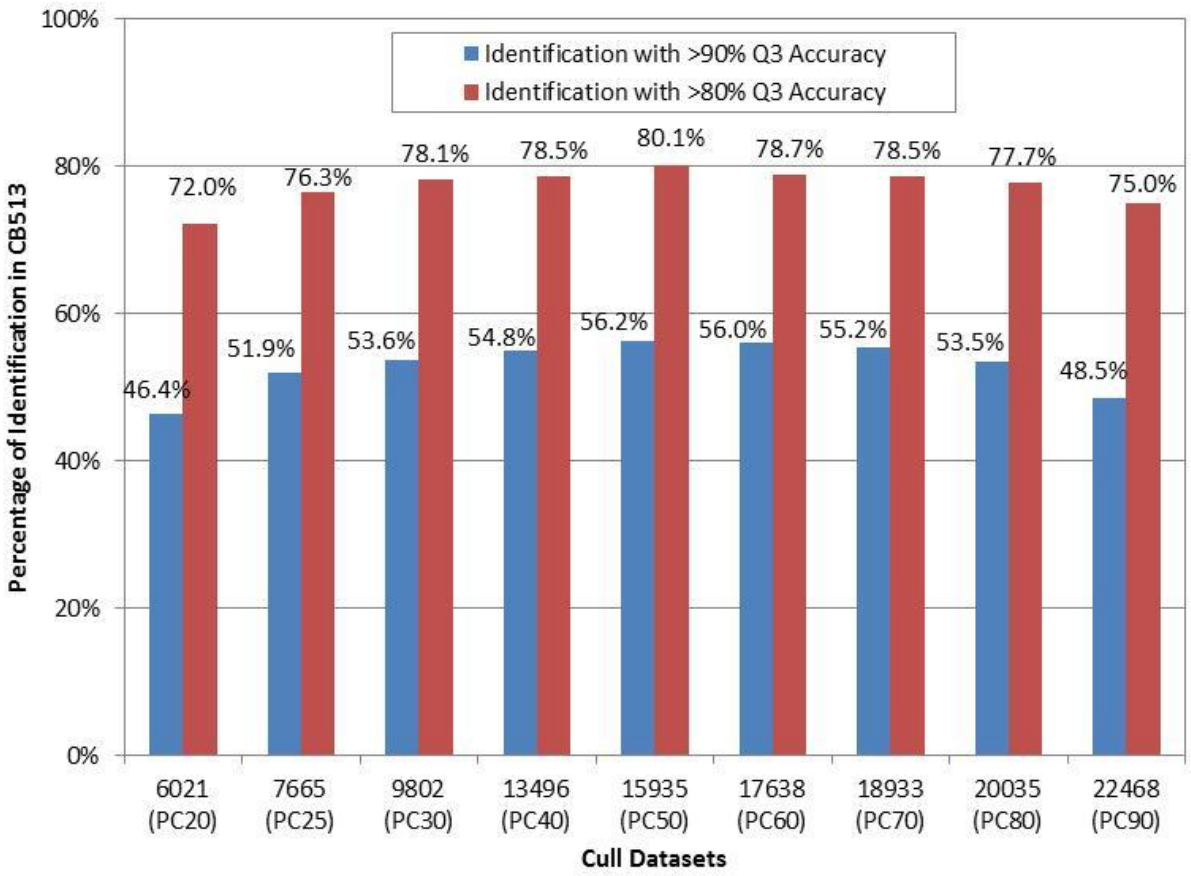


Figure 1. Comparison of identification accuracies of CSSP using different cull datasets with maximum pairwise mutual sequence identity ranging from 20% to 90%. Cull dataset with maximum 50% sequence identity yields best identification accuracy.

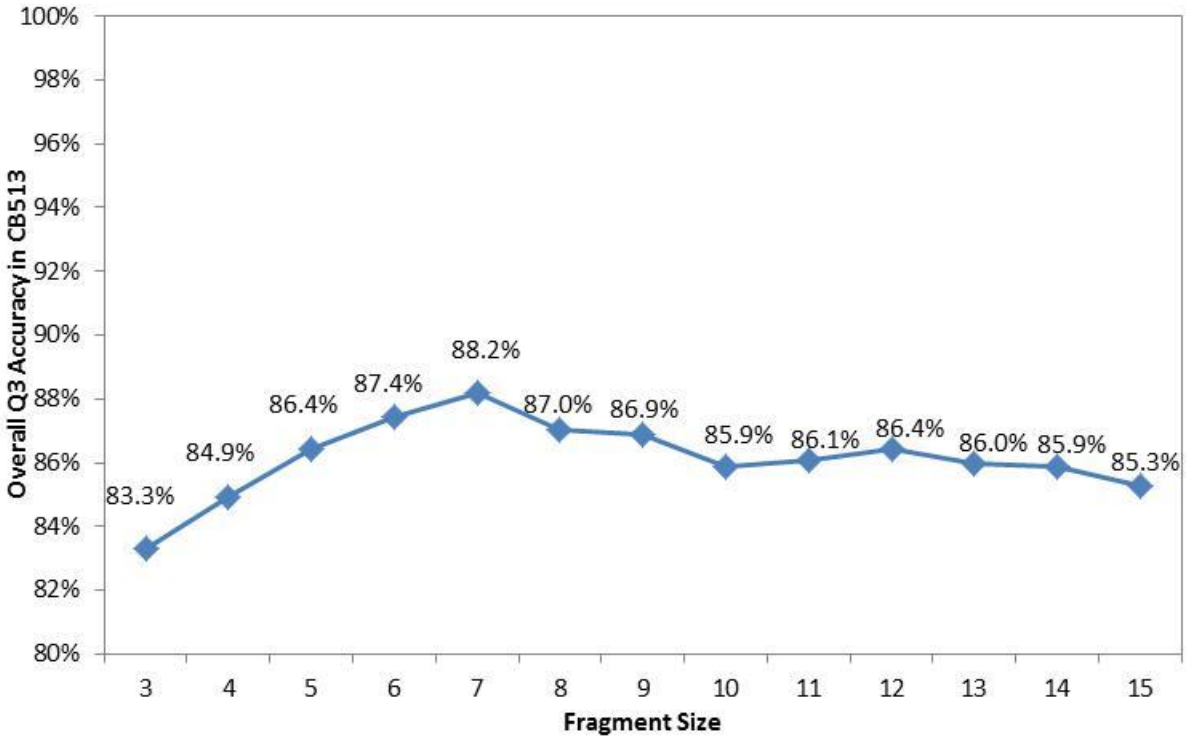


Figure 2. Effect of varying fragment size on the identification accuracy in CSSP using Cull dataset with maximum 50% sequence identity. CSSP with fragment size 7 has the best performance, yielding 88.2% overall Q3 accuracy in CB513.

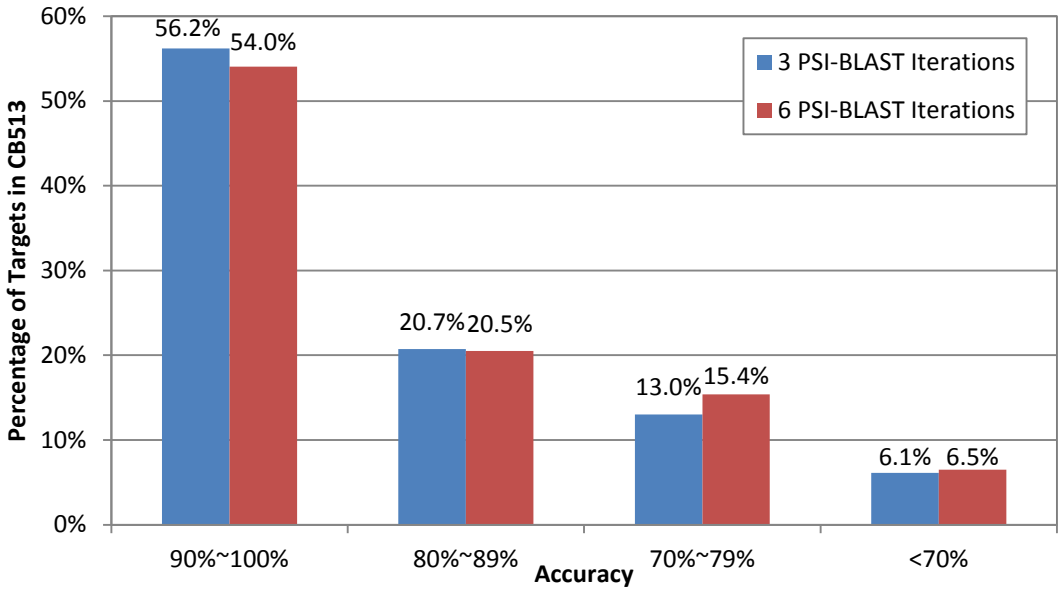


Figure 3. Accuracy Comparison of CSSP using weighted frequencies generated from PSSM using 3 and 6 PSI-BLAST iterations. CSSP based on PSI-BLAST using 3 iterations is slightly more sensitive. Cull dataset with maximum 50% sequence identity and fragment size of 7 are used.

Figure 4 demonstrates the sensitivity of the CSSP potential on 1cdtA in CB513 by comparing the predicted secondary structures by JPred, SAM, and Porter. JPred, SAM, and Porter have Q3 prediction accuracy of 83.3%, 78.6%, and 95.0%, respectively, on 1cdtA. The predicted secondary structure has high Q3 prediction accuracy, which has the similar structure and CSSP potential value as the native. Potential values of each 7-residue fragment in each prediction are displayed in Figure 4. One can notice that mispredicting a β-strand as an α-helix in SAM results in a large spike in the potential values in fragments 35 to 38, indicating that the α-helix is strongly unfavorable. Similarly, the misprediction of a β-strand in JPred leads to significantly higher potential values in fragments from 5 to 21. As a result, Porter's predicted secondary structure has an overall lower potential value (-5.89) than those of JPred (0.28) and SAM (14.64).

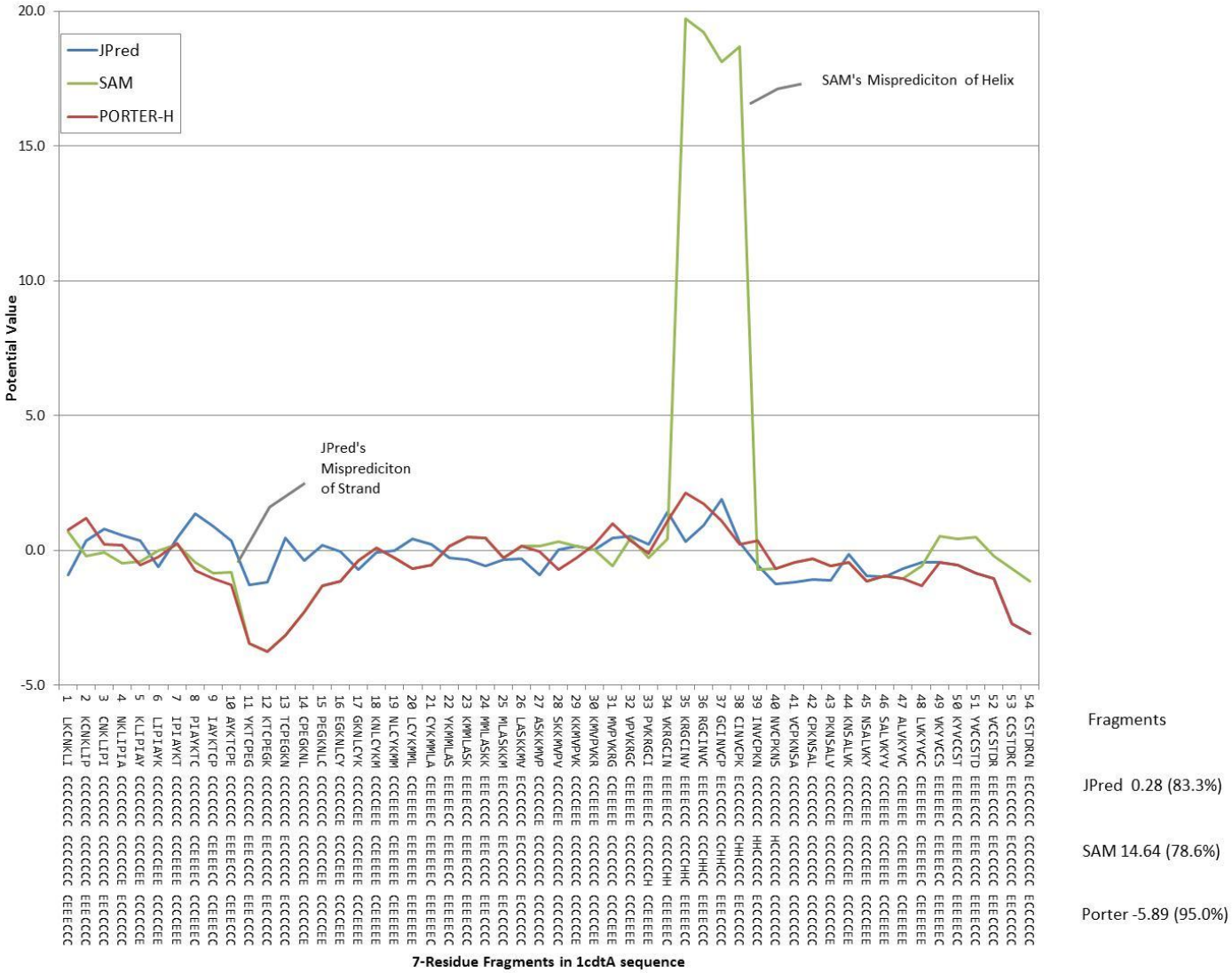


Figure 4. Sensitivity of the knowledge-based potential on 1cdtA. Mispredictions of JPred (fragments 5-21) and SAM (fragments 35-38) lead to higher CSSP potential values than that of Porter's predicted secondary structure.

In more than 80% of the targets in CB513, our best CSSP potential based on triplets picks the predicted secondary structures generated by SSPRO4 or PORTER. This is due to the fact that both SSPRO4 and PORTER take advantage of the structural information of homologues, which is usually helpful to obtain highly accurate prediction. However, when the homologue structures are missing or a wrong homologue template is used, SSPRO4 or PORTER may result in predictions with low accuracy. Figure 5 shows the native secondary structure of 1pyp as well as the predictions

of SSPRO4, PORTER, and PSIPRED. Probably due to lack of homologue structural information in PDB, neither SSPRO4 nor PORTER can reach a prediction with more than 80% Q3 accuracy. In this case, CSSP favors the prediction from PSIPRED, which has the lowest potential value (-6.37) and 81% Q3 accuracy.

```
Sequence (1pyp)
TYTTRQIGAKNTLEYKVYIEKDGKPVSAFHDIPLYADKEDNIFNMVVEIPRWTNAKLEITKEETLNPIIQNTKGKLRFVRNCFPHHG
YIHNYGAFPQTWEDPNVSHPETKAVGDNNPIDVLQIGETIAYTGQVKEVKALGIMALLDEGETDWKVIAIDINDPLAPKLNDIEDVE
KYFPGLLRATDEWFRIYKIPDGKPENQFAFSGEAKNKKYALDIIKETHNSWKQLIAGKSSDSKGIDLTNVTLPDTPTYSKAASDAIP
PASPKADAPIDKSIDKWFF

Native Secondary Structure
CCCEEEECCCCCCCCEEEEECCCECCCCCCCCCCCCCCCCCCECCCCCCCCCECCCCCCCCCCCCCCCCCCCCECCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEECCCCCCCCCCCCECCEEEEEEEEECCCCEEEEEEEECCCCCCCCCCCCHHHH
CCCCCCHHHHHHHHHHHHHHHCCCCCECHHHCCECHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCC

SSPRO4 (Q3: 75%, Potential Value: 0.44)

CEEEEEEEECCCCCCEEEEECCEEECCCCCCCCEEEHHHCEEEEEEEECCCCCECEEECCCCCCCCEECEECCEECECCEECCCCC
CCCEEEEECCCCCCCCCCEECCCCEEECCCCCEEEECCCCCCCCCCCCEEEEEEEEEEEECCCCEEEEEEEECCCCCHHHCCCHHHHH
HHCCCCCHHHHHHHHHCCHHHCCCCCEECHHHCEEEHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCECCCCCCCCECCCHHHHCC
CCEEECCCCCCCCCCCEEC

PORTER (Q3: 72%, Potential Value: 24.50)

CEEEEEEEECCCCCCEEEEECCEEECCCCCCCCEEECCCCEEEEEEEECCCCCEEEEECCCCCCCCCEEECEECEEEECCEECCCCC
CCCEEEEECCCCCCCCCCEECCCCEEECCCCCEEEECCCCCCCCCCCCEEEEEEEEEEEECCEEEEEEEEECCCCCHHHCCCHHHHH
HHCCCHHHHHHHHHHCCHHHCCCCCEEEEEHCEECHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCECCCCCCCCECCHHHHCCC
CCEEECCCCCCHHHHCEEC

PSIPRED (Q3: 81%, Potential Value: -6.37)

CEEEEEECCCCCCCCEEEECCCCCCCCCCCCCCCCCCCCCEEEEEEEECCCCCCEEEEECCCCCCCCCCCCCCCCCCCCEEEEECCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECCCCCCCCCCEEEEEEEEEEECCCCCCCCEEEEEECCCCCCCCCCCCCCHH
HHCHHHHHHHHHHHHCCCCCCCCCCCEEECCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCCCCCEEEEECCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCC
```

Figure 5. Sensitivity of the knowledge-based potential 1pyp. SSPRO4 and PORTER have predictions with Q3 accuracies of 75% and 72%, respectively, due to lack of homologue structural information. Our knowledge-based potential favors the prediction by PSIPRED with Q3 accuracy of 81%.

3.2 CSSP using Doublets, Triplets, and Quartets

Although theoretically, CSSP can incorporate interactions of *k*-lets for arbitrary *k* value, in practice, the accuracy of *k*-let potential is limited by the number of samples available. Figure 6 compares the identification accuracies of CSSP using doublets, triplets, and quartets on CB513 with

fragment size 7 and the cull dataset with maximum 50% sequence identity. One can find that the identification accuracy of CSSP using triplets is significantly higher than CSSP using doublet by incorporating interactions of three residues. Theoretically, CSSP using quartets should have better precision than the one using triplets since higher order of interactions are taken into account. However, as shown in Figure 6, CSSP using quartets is not as accurate as the one using triplets only. Based on the following analysis of sample numbers in doublets, triplets, and quartets, we find that lack of samples in quartets results in significant higher marginal errors in estimating the distribution of secondary structures in quartets than those in triplets. Moreover, CSSP using quartets has almost twice number of terms as CSSP using triplets, which is more prone to suffer from over-fitting, particularly when some terms are under-sampled.

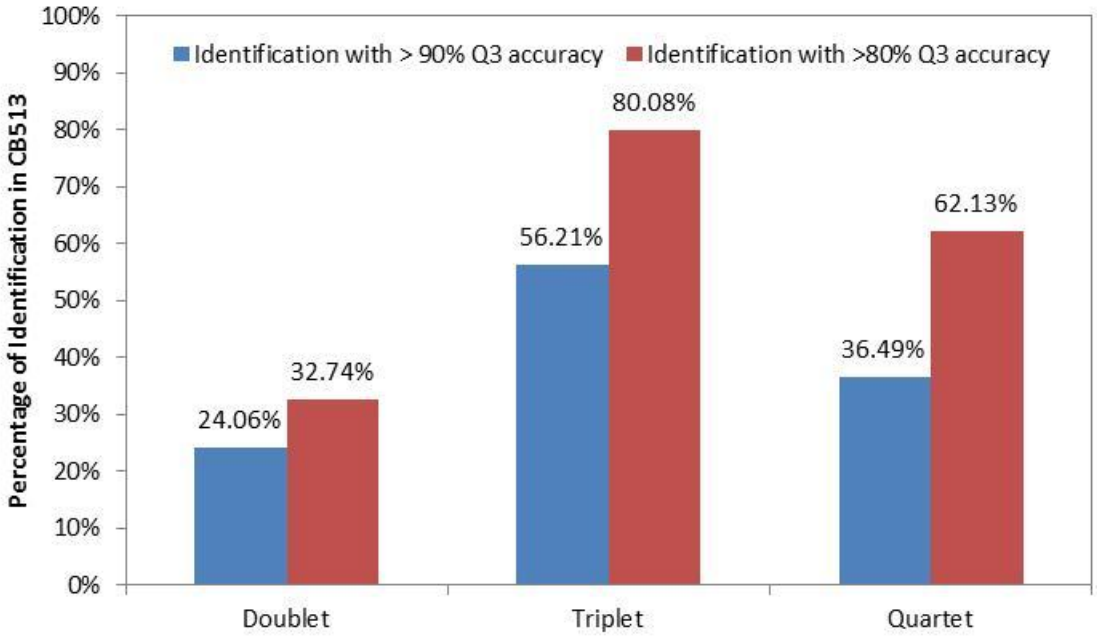


Figure 6. Identification Accuracies of CSSP using Doublets, Triplets, and Quartets in CB513

We use the multinomial distribution to determine the sample size needed to estimate the secondary structure probability of a *k*-let with certain accuracy. Considering statistical samples divided into *m* mutually exclusive and exhaustive categories and denoting π_i, $i = 1, \ldots, m$, to be the

proportion of the samples in the *i*th category. The calculation of the sample size n_i for the *i*th category with precision p_i is

$$n_i = B\pi_i(1-\pi_i)/(1-p_i)^2,$$

where B is the χ^2 value with m - 1 degree of freedom and precision p_i [18]. Let us assume that, in general, the samples are nearly uniformly distributed in the secondary structure categories. The total number of samples needed to estimate the secondary structure distribution of a certain triplet with 99% accuracy (1% marginal error) is

$$n = \frac{45.64\left(\frac{1}{27}\right)\left(1-\frac{1}{27}\right)}{(1-99\%)^2}27 \approx 439496.$$

Similarly, the sample size needed to estimate the secondary structure distribution of a certain quartet with 99% accuracy is

$$n = \frac{112.33\left(\frac{1}{81}\right)\left(1-\frac{1}{81}\right)}{(1-99\%)^2}81 \approx 1109432.$$

	Doublets	Triplets	Quartets
Most number of samples	2731381 (AA @ 0_1)	2416997 (AAA @ 0_1_2)	672129 (AAAA @ 0_1_3_4)
Least number of samples	270313 (WW @ 0_6)	130213 (WWW @ 0_5_6)	2172 (WWCW @ 0_2_3_5)
Average number of samples	1510060	1075624	130565
Percentage with 95% accuracy	100%	100%	85.2%
Percentage with 99% accuracy	100%	94.4%	0%

Table 2. Most, least, and average numbers of samples in doublet, triplets, and quartets at different relative positions when cull dataset with maximum 50% sequence identity is used

Table 2 displays the most, least, and average number of samples in various doublets, triplets, and quartets at different relative positions when the cull dataset with maximum 50% sequence identity is used. Table 2 also shows that 100% and 94.4% of the triplets can achieve 95% and 99% accuracy, respectively. In contrast, only 85.2% of quartets can have 95% accuracy in secondary

structure distribution and none of the quartets achieve 99% accuracy. Particularly for the quartets composed of rare amino acids, the estimated secondary structure distribution has low accuracy. For example, the quartet with minimum samples is WWCW at relative position 0_2_3_5, which has only 2,172 samples – the accuracy of its secondary structure distribution is approximately 80%. As a result, CSSP using quartet is not as precise and sensitive as the one using triplet only. However, the number of high-resolution, experiment-determined protein structures increases rapidly recently. When the protein dataset is grown to about 10 times the size of the dataset we have now, the average accuracies of secondary structure distributions in quartets will reach the current accuracies in triplets and then CSSP using quartets may start to become more effective.

4. Discussion and Summary

In this paper, we present a Context-based Secondary Structure Potential (CSSP) by capturing the high-order inter-residue interactions. The CSSP potential can be effectively used to identify secondary structure predictions with good quality. Moreover, as shown in our computational results and analysis, the CSSP potential using triplets outperforms the CSSP potentials using doublets or quartets. Nevertheless, in the near future when sufficient samples become available, the CSSP potential using quartets may become more effective than the one using triplets.

Although both CSSP and GOR5[27] explicitly consider the high-order inter-residue interactions, the mechanisms of calculating and integrating these interactions are different due to different purposes of CSSP and GOR5. The goal of GOR5 is to predict the secondary structure of each residue. Therefore, the GOR5 scores evaluate how likely a residue adopts a certain secondary structure within its amino acid environment. However, GOR5 is unable to take the influence to a residue from the secondary structures of its neighboring residues into account because they are unknown. In fact, the secondary structures of the neighboring residues play an important role. For

example, if the adjacent positions of a residue are not helices, it is impossible for this middle residue to adopt helix as its secondary structure. In contrast, the purpose of CSSP is to assess the qualities of predicted secondary structures, where the favorability of two, three, four, and theoretically up to *k* residues concurrently adopting certain secondary structures are of interest. Compared to the secondary structure energy by Miyazawa and Jernigan[29], CSSP considers more general *N*-body interactions among not necessarily consecutive residues. CSSP is also different from the *k*-mer model[28] proposed by Madera et al., whose purpose is to refine secondary structure predictions and the *k*-mer contains the secondary structure information only. In comparison, the *k*-lets in CSSP measure the high-order correlation between sequence and structure, which include both sequence and structure information.

One of the main disadvantages of the CSSP potential capturing high-order inter-residue interactions is its high computational cost. Considering the CSSP potential with fragment size *N* and calculating up to *k*-let inter-residue interactions, the total number of *k*-let calculations for a protein with *P* residues is

$$P/N \sum_{i=1}^{k} \binom{i}{N}.$$

In our computational experiments, calculating CSSP potential using triplets for one postulated protein structure takes a few seconds to several minutes on a single processor. CSSP potential using quartets is even more computationally cost. Therefore, we use CSSP to assess predicted secondary structures instead of using CSSP to predict secondary structures. Nevertheless, computing CSSP potential is data-intensive and parallelizable. Taking advantage of the emerging massively parallel computing architectures such as Graphics Process Units (GPU) and data-intensive parallel computing algorithms[19], one may be able to reduce the computational time of evaluating CSSP significantly and then use CSSP efficiently for secondary structure prediction. Another disadvantage is that the current CSSP is unable to capture global interactions exceeding the fragment size.

Acknowledgements

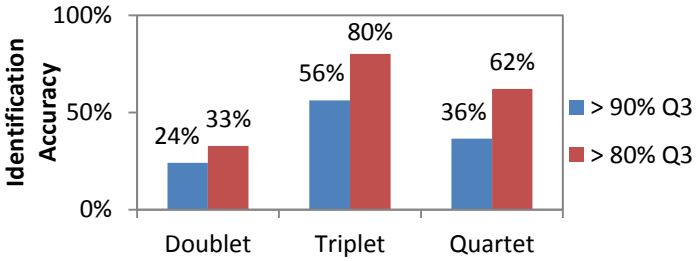
YL acknowledges support from NSF under grant 1066471 and ODU 2011 SEECR grant. IR acknowledges support from CNCSIS-UEFISCDI under project number PN-II-PT-PCCA-2011-3.1-1350.

References

(1) M. J. Sippl, "Calculation of conformational ensembles from potentials of mean force – an approach to the knowledge-based prediction of local structures in globular proteins." J. Mol. Biol., **213**: 859-883, 1990.
(2) R. Samudrala, J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction," J. Mol. Biol., **275**: 895-916, 1998.
(3) G. Wang, R. L. J. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, **19**:1589-1591, 2003.
(4) S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped Blast and Psi-Blast: A New Generation of Protein Database Search Programs," Nucleic Acids Research, **25**: 3389-3402, 1997.
(5) J. A. Cuff, G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," Proteins, **34**: 508-519, 1999.
(6) C. Mooney, G. Pollastri. "Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information," Proteins, **77**(1): 181-90, 2009.
(7) J. Garnier, J. F. Gibrat, B. Robson, "GOR method for predicting protein secondary structure from amino acid sequence," Methods Enzymol, **266**: 540-553, 1996.
(8) Y. Guermeur, "Combinaison de classifieurs statistiques, application a la prediction de la structure secondaire des proteins," Ph.D. Thesis, 1997.
(9) K. Karplus, "SAM-T08, HMM-based protein structure prediction," Nucleic Acids Research, **37**(2): W492-497, 2009.
(10) C. Cole, J. D. Barber, G. J. Barton, "The Jpred 3 secondary structure prediction server," Nucleic Acids Research, **36**(2): W197-201, 2008.
(11) M. Ouali, R. D. King, "Cascaded multiple classifiers for secondary structure prediction," Protein Science, **9**(6): 1162-1176, 2000.
(12) D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," J. Mol. Biol., **292**: 195-202, 1999.
(13) B. Rost, C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," Proteins, **19**(1): 55-72, 1994.

(14) J. Meiler, D. Baker, “Coupled prediction of protein secondary and tertiary structure,” Proc. Natl. Acad. Sci., **100***:* 12105–12110, 2003.
(15) B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, C. Lundegaard, “A generic method for assignment of reliability scores applied to solvent accessibility predictions,” BMC Structural biology, **9**(1): 51, 2009.
(16) J. Cheng, A. Randall, M. Sweredoski, P. Baldi, “SCRATCH: a Protein Structure and Structural Feature Prediction Server,” Nucleic Acids Research, **33**: 72-76, 2005.
(17) G. Pollastri, A. McLysaght, “Porter: a new, accurate server for protein secondary structure prediction,” Bioinformatics, **21**(8): 1719-1720, 2005.
(18) R. Tortora, “A note on sample size estimation for multinomial populations,” The American Statistician, **32**(3): 100-102, 1978.
(19) A. Yaseen, Y. Li, “Accelerating Knowledge-based Energy Evaluation in Protein Structure Modeling with Graphics Processing Units,” Journal of Parallel and Distributed Computing, **72**(2): 297-307, 2012.
(20) D. Kihara, “The effect of long-range interactions on the secondary structure formation of proteins,” Protein Science, **14**(8): 1955-1963, 2005.
(21) B. Rost, “Review: protein secondary structure prediction continues to rise,” Journal of Structural Biology, **134**(2-3): 204-218, 2001.
(22) W. Kabsch, C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” Biopolymers, **22**(12): 2577-2637, 1983.
(23) D. Frishman, P. Argos, “Knowledge-based protein secondary structure assignment,” Proteins, **23**(4): 566-579, 1995.
(24) O. Dor, Y. Zhou, “Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training,” Proteins, **66** (4): 838–45, 2006.
(25) W. Pirovano, J. Heringa, "Protein Secondary Structure Prediction," Methods in Molecular Biology, **609**(3): 327-348, 2010.
(26) Y. Wei, J. Thompson, C. A. Floudas, "CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization," Proc. R. Sco. A, **468**: 831-850, 2012.
(27) A. Kloczkowski, K. L. Ting, R. L. Jernigan, J. Garnier, "Combining the GORV Algorithm With Evolutionary Information for Protein Secondary Structure Prediction From Amino Acid Sequence," Proteins, **49**: 154–166, 2002.
(28) M. Madera, R. Calmus, G. Thiltgen, K. Karplus, J. Gough, "Improving protein secondary structure prediction using a simple *k*-mer model," Bioinformatics, **26**(5): 596-602, 2010.
(29) S. Miyazawa, R. L. Jernigan, "Evaluation of short range interactions as secondary structure energies for protein fold and sequence recognition," Proteins, **36**: 347-356, 1999.

Table of Contents Graphic



Implementarea metodelor si protocoalelor de testare pe soareci de laborator a interactiei de tip EPI a bacteriilor cu molecule de medicamente

UMF „Carol Davila"/Oftalmologie Clinica

Infectiile intraoculare reprezinta un punct de cotitura pentru oricare oftalmolog, fiind extrem de dificil de tratat. Ochiul ca sistem inchis, in majoritatea cazurilor nu poate drena sau asana o infectie intraoculara, motiv pentru care chiar si atunci cand tratamentul este instituit de timpuriu, poate fi foarte greu de salvat ochiul si sunctia vizuala. Endoftalmita este una din cele mai devastatoare complicatii oculare ce poat aparea dupa chirurgia oculara, sau dupa patrunderea in cavitatea oculara a unor corpi straini. Este esentiala diagnosticarea cat mai precoce pentru pastrarea ochiului ca organ si a vederii.

Tratamentul unei endoftalmite este chiar si in ziua de azi o provocare. In general afectiunea este simptomatica (dureri oculare, scaderea acuitatii vizuale, hiperemie conjunctivala si chemozis conjunctival, edem palpebral, edem cornean, hipopion sau fibrina in camera anterioara, eventual defect pupilar aferent relativ. Segmentul posterior este greu de vizualizat, dar atunci cand este posibil se observa un vitros tulbure, cu abcese vitreene si intecuiri vasculare. In unele cazuri, afectiuena poate fi asimptomatica sau cu semne extrem de vagi.

In prezent, tratamentul endoftalmitelor nu a cunoscut o evolutie spectaculoasa. Initierea unui tratament cu corticosterozi poate avea un rezultat pozitiv intr-o prima faza; diagnosticul diferential intre un sindrom inflamator si endoftalmita se face prin biopsie vitreana sau prelevare de umor apos din camera anterioara.

Pana in prezent, este cunoscut un singur studiu randomizat prospectiv privind managementul endoftalmitelor (The Endophthalmitis Vitrectomy Study - EVS), desfasurat in Statele Unite, care a simplificat modul de abordare al tratamentului endoftalmitelor, insistand pe importanta instalarii terapiei din momentul punerii diagnosticului.

In toate cazurile in care acuitatea vizuala este mai buna de perceperea luminii, se recomanda efectuarea unei biopsii printr- vitrectomie prin pars plana. Din specimenele obtinute se efectueaza culturi pentru identificarea germenilor si testarea responsivitatii lor la antibiotice. Spatiul creat in cavitatea vitreana prin biopsie poate fi folosit pentru injectarea de antibiotice. In studiul EVS au fost utilizate amikacina si vancomicina. Gentamicina si Cefuroximul au aproximativ acelasi spectru de actiune. Acelasi studiu a aratat ca antibioticele injectate intravenos nu aduc niciun beneficiu.

In ce priveste vitrectomia cu triplu abord prin pars plana,ea este indicata doar cand acuitatea vizuala este de la percepe lumina in sus. Antibioticele topice administrate intensiv nu sunt in mod special indicate, exceptand situatia unei plagi oculare cu probleme specifice sau a unei keratite microbiene.

EVS nu a investigat beneficiile oferite de steroizii administrati intravitrean, dar pana in prezent folosirea lor nu are baza.prednisolonul administrat sistemic in doze mari poate fi administrat 60-80 mg/zi, scazand dozele treptat in 7-10 zile. In cazul in care vorbim de o endoftalmita fungica, streoizii sunt contraindicati in endoftalmitele cu etiologie fungica.

Administarea intravitreala a medicamentelor

Doza de antibiotic administrata intravitrean este cuprinsa in 0.1 ml de substanta; cand vorbim de combinatii intre mai multe antibiotice, atunci doza administrata este de 0.2 ml. Sunt indicate seringi speciale de 1 ml (tip insulina), cu ac de 25 sau 27 gauge.

1. Gentamicina

Doza necesara administrarii intravitreene: 200µg in 0.1ml;

*. Se extrag 0.5ml dintr-un flacon cu gentamicina ce contine 40mg/ml;

*. Se completeaza pana la 10ml cu solutie salina sau BSS (balanced salt solution);

*. 0.1ml din aceasta solutie va contine 200µg de antibiotic.

2. Amikacina

Doza necesara administrarii intravitreene: 0.4mg in 0.1ml;

*. O fiola de antibiotic(500mg) se extrage intr-o seringa de 10 ml si se completeaza cu BSS;

*. Din solutia preparata se extrag 0.8ml (folosind o seringa speciala de 1ml) si se completeaza cu BSS

pana la 10 ml;
*. Din aceasta solutie se extrag 0.1ml continand 0.4mg de substanta.
3. Cefuroxim sau Vancomicina
Doza necesara administrarii intravitreene: 1000µg in 0.1ml;
*. Fiola de 250mg de antibiotic se completeaza cu 8ml de solutie salina sau BSS;
*. Acest amestec se completeaza pana la 10 ml solutie salina sau BSS;
*. Se injecteaza 2ml inapoi in fiola si se completeaza pana la 5ml BSS sau solutie salina;
*. 0.1ml din aceasta solutie reprezinta 1mg (1000µg)
4. Amfotericina
Doza necesara administrarii intravitreene:5µg in 0.1ml
*. Se completeaza fiola de 50 mg de substanta cu BSS/solutie salina, pana la 10 ml in total;
*. Din acest amestec se extrag 0.1ml si se completeaza pana la 10 ml cu BSS sau aolutie salina; 0
*. 0.1ml din acest amestec contin 5µg de substanta;
Alternativ, se poate injecta fiola de 50 mg intr-o punga de 1l de solutie Ringer. 0.1 ml din acest amestec contine 5µg de substanta.
5. Clindamicina
Doza necesara administrarii intravitreene: 1000µg in 0.1ml
*. Se extrage continutul unei fiole (2 ml=300mg) si se completeaza pana la 3 ml cu BSS sau solutie salina;
*. Din acest amestec se retrag 1 ml si se completeaza pana la 10 ml de BSS/solutie salina;
*. 0.1ml din acest amestec contine 1000µg
Indicatie bibliografica: http://www.mrcophth.com/focus1/endophthalmitis.html

Rezistenta la tratamente multiple (MDR) reprezinta o problema din ce in ce mai des intalnita in oftalmologie, reprezentand o problema majora in managementul infectiilor oculare si sistemice, in special daca vorbim de endoftalmitele cu acesti germeni rezistenti la tratamentele uzuale. Rezistenta la tratamente multiple (MDR) se refera la rezistenta unui germene la 2 sau mai multe clase de antibiotice. In studiul EVS este aratat faptul ca 100% din bacteriile Gram pozitive raspund la Vancomicina, 89% din bacteriile Gram negative sunt susceptibile la Ceftazidima si Amikacina, in timp ce restul de 11% sunt rezistente la aceste antibiotice. Studii retrospective efectuate in India in perioada 2000-2007 releva faptul ca bacteriile rezistente sunt in majoritatea cazurilor Gram negativ (speciile de Pseudomonas) si, conforma acestui studiu, prognosticul este unul prost, aceste bacterii fiind rezistente la Vancomicina si Amikacina.

Pseudomonas aeruginosa dobandeste rezistenta la antibiotice prin posesia unor gene rezistente codificate la nivel cromozomial, ce produc excesiv AmpC cefalosporinaza, ce confera bacteriei rezistenta la toate beta lactaminele, exceptand carbapenemii. Rezistenta la carbapenemi se produce prin down-regulare la nivelul proteinei membranare externe (OprD) care reprezinta calea primara de patrundere a carbapenemilor. Pompele de eflux au abilitatea de a exclude multe antibiotice de la nivel periplasmic sau citoplasmatic. Exprimarea naturala a pompelor de eflux au un rol important in susceptibilitatea relativ scazuta a Pseudomonas aeruginosa la antibiotice.

In ceea ce priveste bacteriile Gram pozitive, intre cele mair ezistente se numara bacteriile din grupul Enterococcus, rezistente in special la Vancomicina, si care in unele cazuri pot duce la aparitia phtisis bulbi.

In urma studiului efectuat in India, s-a constatat ca atat unele bacterii Gram pozitive cat si unele Gram negative sunt rezistente la Gatifloxacin (fluoroquinolona de generatia a IV-a). Acest fapt este cu atat mai alarmant cu cat fuoroquinolonele de generatia a IV-a inhiba topoizomerazele II si IV, fapt care ar fi facut mai putin probabila dobandirea unei rezistente bacteriene.

Tot acest studiu a aratat ca 71% din pacientii cu endoftalmita cu germeni rezistenti la tratament au un prognostic vizual prost. O posibila cale de a combate aceasta rezistenta ar fi cresterea compliantei la tratamentele de lunga durata cu antibiotice sau evitarea folosirii antibioticelor cu spectru larg pentru afectiuni ce nu indica acest lucru.
Indicatie bibliogtrafica: http://group.irso.org/knowing/5.pdf

In esenta studiile noastre au aratat ca: Rezistanta multipla dezvoltata de bacterii si tumori poate fi

invinsa prin modificarea structurii moleculare a medicamentelor in urma expunerii acestora la radiatie laser; Clorpromazina (CPZ) expusa la radiatie laser devine eficienta impotriva unor culturi de *Staphilococcus aureus*; se observa un posibil tratament mai efficient al pseudotumorilor produse in ochi de iepure prin intermediul utilizarii medicamentelor irradiate cu radiatie laser; medicamentele utilizate au fost: Clorpromazina (CPZ) in apa distilata (10mg/ml and 20 mg/ml); s-au folosit atat probe neiradiate cat si iradiate cu un laser Nd:YAG (λ=266nm, E=6.5mJ), intr-un intervat de timp cuprins intre 5 minute si 4 ore; masuratorile au aratat modificari ale spectrelor de absorbtie ale CPZ iradiat, corelate cu timpul de iradiere, indicand modificarile induse in moleculele de CPZ (Fig.12).

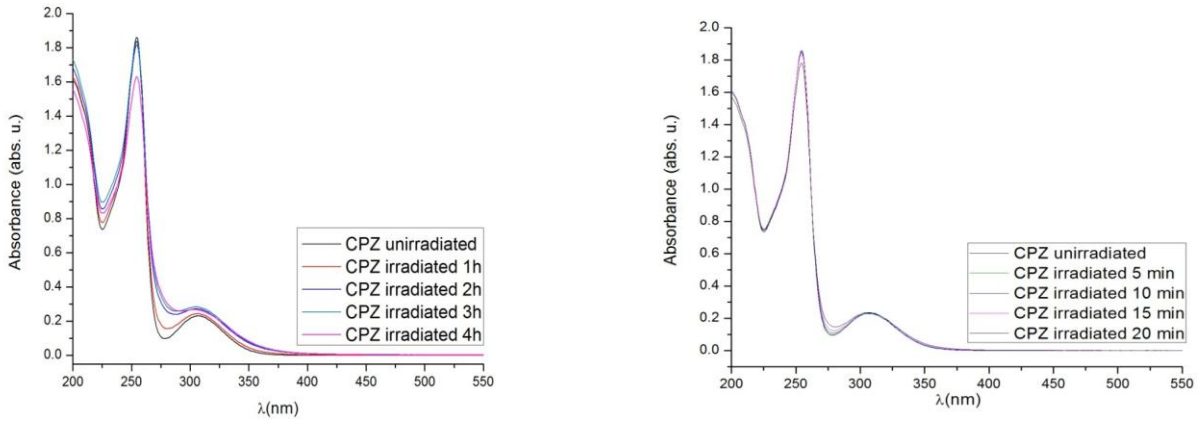


Fig.12 a). Timpi de iradiere pana la 4 ore; spectre de absorbtie masurate intre 200 nm si 550 nm.
b).Timpi de iradiere pana la 20 minute; spectre de absrobtie masurate intre 200 nm si 550 nm**.**

Experimentele pe ochi de animal au constat in : *Modele experimentale; 5 iepuri, cu varste intre 8 luni si un an; *Au fost produse pseudotumori la limbul sclero-corneal utilizand propilen 5.0 (Fig.13A,B);

Fig. 13 A si B.Aspectul pseudotumorilor obtinute;

*Dupa 7 zile s-au tratat ochii dupa cum urmeaza:^**Primul iepure** –ambii ochi au fost netratati si pastrati ci masuratoare de control (aspect histologic - Fig.14);^**Al doilea iepure –** au fost tratati ambii ochi prin injectarea in pseudotumori astfel: primul ochi 0.1 ml CPZ 20 mg/ml neiradiat (Fig. 15A), al doilea ochi 0.1ml CPZ 10 mg/ml neiradiat (Fig. 15B);^**Al treilea si al patrulea iepure –** au fost tratati ambii ochi primul ochi al fiecarui iepure cu 0.1ml CPZ 20 mg/ml iradiat 20 minute (Fig. 16A), al doilea ochi cu 0.1 ml CPZ 10 mg/ml iradiat 20 minutes (Fig. 16B);^**Al cincilea iepure** a fost tratat in primul ochi

injectandu-se cu 0.1ml CPZ 20 mg/ml iradiat 4 ore (Fig. 17A) iar al doilea ochi cu 0.1 ml CPZ 10 mg/ml iradiat 4 ore (Fig. 17B);*3 zile dupa injectarea substantelor s-au efectuat masuratori anatomo-patologice.

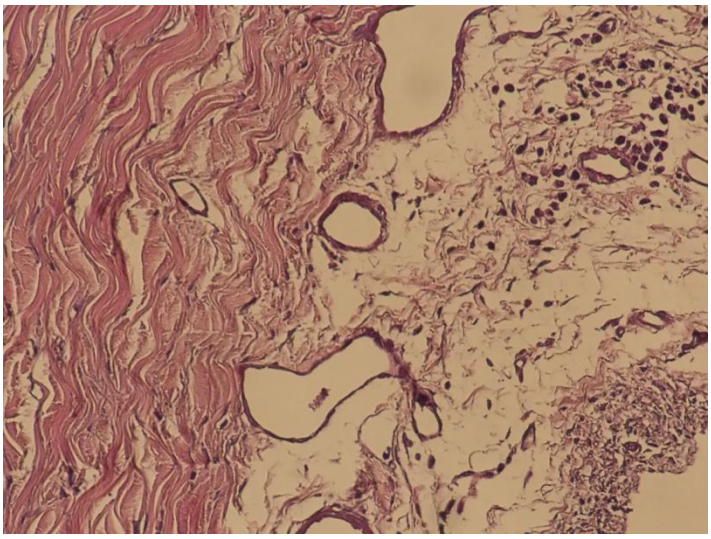


Fig. 14 Tesut inflamator, multe neovase, fibroza locala

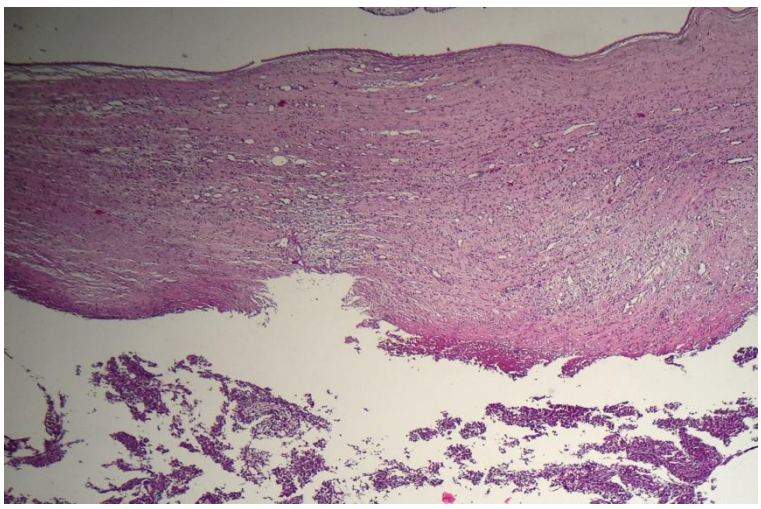


Fig. 15A Cantitate mare de tesut inflamator, multe neovase, fibroza locala, necroza a tesutului

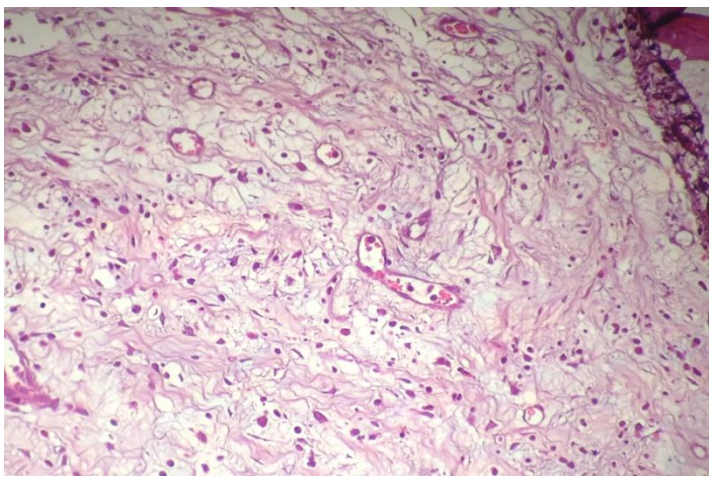


Fig. 15B Tesut inflamator neovase, fibroza locala, mai putine eozinofile decat in fig 16A

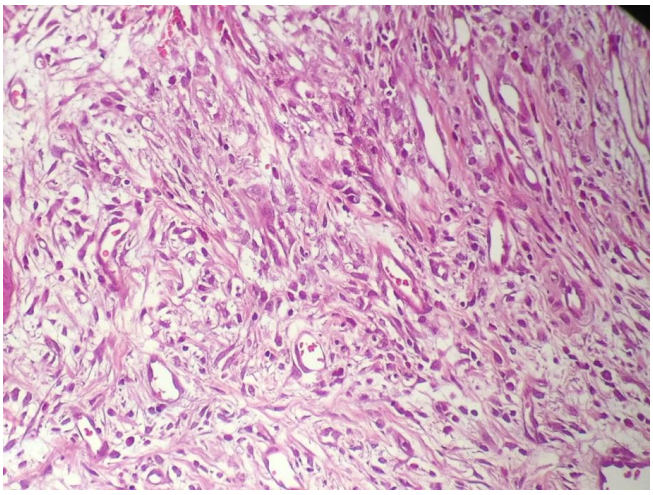


Fig. 16A Tesut inflamator, neovase, fibroza locala

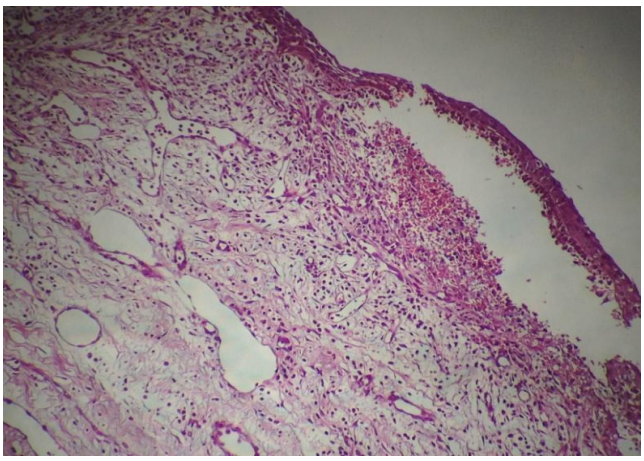


Fig. 16B Mai putin tesut inflamator decat in cazul fig 15, neovase, fibroza locala, eozinofile absente, tesut fibrotic mai ridicat

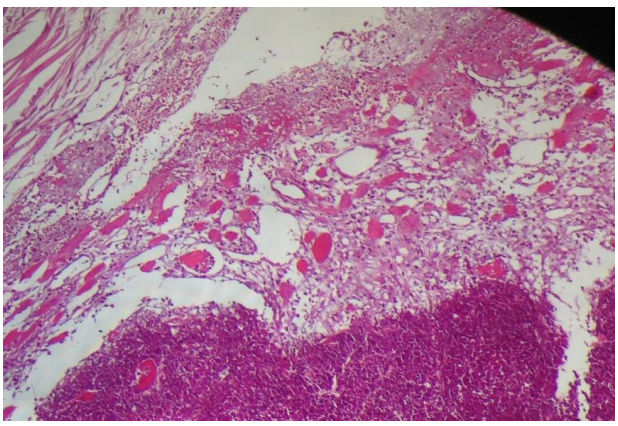


Fig. 17A Tesut inflamator, neovase, fibroza locala

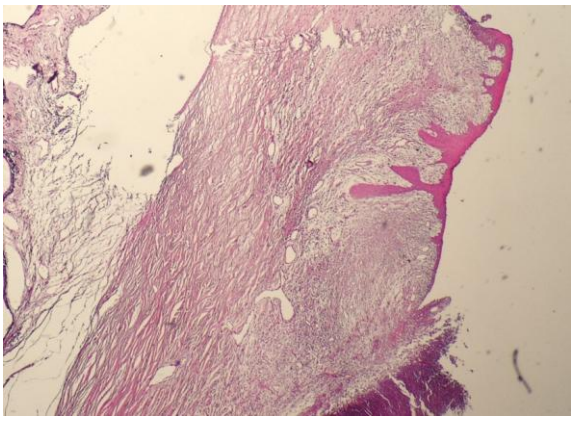


Fig. 17B Tesut inflamator, (mai putin decat in fig 16B), neovase, fibroza locala

Concluziile acestor studii preliminare sunt: * Utilizarea CPZ iradiat pentru tratarea tesuturilor pseudotumorale produce efecte care depind de concentratia CPZ in solvent si de timpul de iradiere a solutiilor ; *Moleculele de CPZ expuse la radiatie laser si modificate de aceasta pot devein eficiente in tratarea tesututilor pseudotumorale si posibil a tumorilor maligne care au dezvoltat MDR.