РЕЦЕНЗИЯ

от проф. д-р Никола Ив. Янев

на докторска дисертация на тема: Моделиране и оптимизация при медико-биологични изследвания

Автор: РАДОСЛАВ СТЕФАНОВ МАВРЕВСКИ

Област на висше образование 4. Природни науки, математика и информатика.Професионално направление 4.6. Информатика и компютърни науки.Докторска програма "Информатика"

Резултатите в дисертацията (140 страници -4 глави,7 приложения, 160 цитирания) са в биоинформатиката: интердисциплинарна област, развиваща методи и софтуерни средства за тълкуване на биологични данни. Като интердисциплинарна област на науката, биоинформатиката комбинира компютърни науки,статистика, математика и инженерство за изучаване и обработка на биологични данни. В смисъла на тази дефиниция дисертацията е в указаната в4.6 област.
Основен резултат: **създаване на относително интегрирана среда включваща методология и софтуер за използване в медико-биологични изследвания на развити програмни средства като: Matlab,GOLD,Molegro Virtual Docker,SPSS Statistics, GraphPad PRISM**.
Необходимостта от това е обусловена от съвременната представа за медико-биологичните изследвания,които

са свързани с проектирането на биологични експерименти, събиране, обобщаване и анализ на данни от тези експерименти, както и тълкуването на резултатите и изводите от тях. Статистическите методи, математическото и компютърното моделиране са от критично значение за медико-биологичните изследвания. Обединяването на целта на моделирането и експериментални

изследвания при медико-биологичните науки е свързано с изясняване на основните биологични процеси, в резултат на дадени наблюдавани явления. Важно предизвикателство за математическото моделиране включва определяне на подходящ модел от даден клас, който да бъде използван за да се правят заключения и прогнози..

Необходимостта от математическо и компютърно моделиране в медико-биологичните изследвания се обуславя от факта, че непосредственото изследване на много обекти е практически невъзможно или изисква много време, средства и специализирано техническо оборудване за експерименти.

Перспективите за развитие на математически модели в медико-биологичните изследвания почиват на използването на информационните технологии. Информационните технологии позволяват интегрирането на знания под формата на математически обекти и под формата на визуални изображения, което представя идеята на сложни закони на функциониране в живи системи, които е трудно да бъдат формализирани. Биоинформатиката използва компютрите за събиране, съхраняване, анализиране и интегриране на биологична и генетична информация, която след това може да се приложи например, при откриване и разработване на лекарства. Целите на биоинформатиката са: организиране на данни по начин, който позволява на учените да имат достъп до наличната информация и да представят нови записи; да се разработят инструменти и ресурси, които помагат при анализа на данните; да се използват тези инструменти, за да се анализират данните и интерпретират резултатите по биологично значим начин.

Математическо и компютърното моделиране са неразделна част от научните изследвания в много области.. Съществен момент при математическия модел

е степента на точност, респективно адекватността с реалния обект, която е в тясна зависимост с парадигмите в конкретната научна област (Mason R. L., R. F. Gunst, J. L. Hess, 2003). Проблемът за избор на „оптимален“ модел е един от най-основните проблеми в анализа на данни. Изследването за избор на „оптимален“ математически модел е един актуален научен проблем, което е видно от големия брои световно извести учени, които работят по този проблем.

Приложната част на дисертацията е конкретизация на развити техники за моделиране, статистически анализ , избор на оптимален модел върху данни от експерименти провеждани във Факултета по обществено здраве и спорт на Югозападния Университет “Неофит Рилски” в Благоевград. Предмет на изследването е избора и оценката на модели, използвани в медико-биологични изследвания като: зависимостта торг-ъглова позиция на екстензорите на лакътя; максимална кислородна консумация VO_{2max}; и зависимост между резултати от докинг и биологични изследвания при лиганд-рецепторни взаимодействия.

За намиране на индивидуалните „оптимални“ модели в класовете кандидат модели при моделиране на зависимостта торг-ъглова позиция са използвани различни регресионни (fitting) методи, като най-широко използваният метод на най-малките квадрати (LS), стабилна регресия (RR) и минимакс fitting. В случай на LS и RR е използван софтуерния продукт GraphPad Prism 6.0. В случай на минимакс fitting е използван Matlab и специална функция fminmax, като за целта са създадени Matlab скриптове за минимаксна оптимизация. Наборът от „оптимални“ модели от различни класове е сравнен, с разгледаните съществуващи и ново предложените в дисертацията различни критерии за оценка качеството на тези модели.

Такива често използвани в литературата, критерии за оценката на модели от различни класове са например информационен критерий на Акайке (AIC), информационен критерий на Бейс (BIC) и средно квадратично отклонение (RMSD). В информационно-теоретичния подход, препоръчван от (Akaike H., 1974; Kullback S., R. A. Leibler, 1951), информационното несъответствието се счита като основен критерий за оценка на качеството на модел като приближение до истинското разпределение, което генерира данните. С разработването на гъвкави техники за моделиране, става необходимо разработването на критерии за избор на модели от различни класове, оценявани по методи, различни от метода на максималната вероятност. В дисертацията са предложени и минимизиране на максималния остатък (MMR) и Хаусдорфовото разстояние критерий (HDC), като критерии за избор на модел и са сравнени с други известни в литературата критерии. Разработена е програмата „Comparing models" изчисляваща критериите за избор на „оптимален" модел (AIC, BIC, RMSD, MMR и HDC), реализирана като отделни модули, всеки от които служи за пресмятането на съответния критерий.

За оценка на параметрите в уравнението за изчисляване на максимална кислородна консумация VO_{2max} за изследваните лица, е направен многофакторен регресионен анализ по метод на най-малките квадрати със софтуерния продукт SPSS Statistics. Разработена е също и софтуерна програма за пресмятане на VO_{2max} с тест на Astrand-Ryhming.

При моделиране на лиганд-рецепторните взаимодействия са използвани софтуерните продукти GOLD и Molegro Virtual Docker за определяне на силата и начина на свързване на лиганда с рецептора и софтуерния продукт Avogadro за моделиране и оптимизиране на структурите на лигандите.

Най-общо, резултатите са:

- моделиране на торг-ъглова позиция при биомеханични изследвания;

- моделиране с регресионно уравнение за определяне на $VO_{2max;}$

- моделиране на лиганд-рецепторни взаимодействия.

Към получените резултати следва да се включи и втора глава, съдържаща описание на използваните методични подходи при медико-биологични изследвания и моделирането на зависимости. Представена е методологията на изследванията при отделните обекти, статистическата обработка на експерименталните резултати, математическото и компютърното моделиране на зависимостите и използваните критерии за избор на „оптимален“ модел. Това е всъщност оригинален обзор върху десетки статии и книги, третиращи горната тематика и който сравнително лесно може да бъде превърнат в цикъл лекции за обучение на студенти по биоинформатика или поне в полезно ръководство за биолози занимаващи се с подобни изследвания, каквато е например книгата:

GraphPad PRISM
Fitting Models to Biological data using Linear and Nonlinear Regression
A practical guide to curve fitting
Harvey Motulsky , Arthur Christopoulus

По-детайлни резултати:
1. Предложена е една последователна стратегия при анализа и моделирането на експериментални данни при медико-биологични изследвания;

2. Предложен е „оптимален“ математически модел на торг-ъглова позиция зависимости в областта на биомеханиката, базирайки се на различни

фитващи методи като най-малките квадрати, стабилна регресия и минимкас фитване, както и подходящ софтуер за регресионен анализ;

3. Приложени са често използвани критерии AIC, BIC и RMSD за избор на „оптимален“ модел от различни класове модели описващ торг-ъглова позиция зависимости;

4. Предложени също два нови критерия MMR и HDC и е оценена тяхната ефективност като такива, чрез сравняването им с AIC, BIC и RMSD;

5. Разработена е софтуерна програма за пресмятане на критериите AIC, BIC, RMSD, MMR и HDC за избор на „оптимален“ модел от различни класове модели;

6. Намереният „оптимален“ модел подпомага биомеханичния анализ на поведението на екстензорите на лакътя за генериране на максимална изометрична сила съобразно изходната дължина на мускулите;

7. В резултат от направеният многофакторен регресионен анализ с SPSS за оценка на параметрите в регресионното уравнение за пресмятане на VO_{2max} са намерени „оптималните“ параметри в модела при нетренирани лица;

8. Разработена е софтуерна програма за пресмятане на VO_{2max} с тест на Astrand-Ryhming, която улеснява определянето на VO_{2max} с този индиректен тест;

9. Установена е голяма корелация (r = -0.72) между данните получени от докинг изследвания с GOLD и *in vitro* измервания; това доказва надеждността на резултатите при използване на GOLD и ДОР при сравнително изследване с MVD.

Получените резултати са публикувани в *3* статии в списания (от които *1* самостоятелна в списание Доклади на БАН), и *4* в доклади (от които *1* самостоятелна) на международни конференции. *1* статия е подготвена за публикуване . *7* публикации са на английски език и *1* е на български език.

Обучението на докторанта включва и участие в интензивни курсове:

- DAAD Intensive course "Symmetry in Science and Art", 10-16 May, 2011, Vrnjacka Banja, SERBIA;
- DAAD Intensive course "Applications of the Calculus of Variations and Optimal Control. The Smooth and Nonsmooth Cases", 11-18 July, 2011, Struga, MACEDONIA;
- DAAD Intensive course "Robotics and Mathematics", 12-18 August, 2012, Ohrid, MACEDONIA;
- DAAD Intensive course " Summability Theory and Statistical Convergences", 20-27 August, 2012, Pristina, KOSOVO;
- Курс Суперкомпютърни приложения в природните науки, „Използване на системната и приложната среда на Blue. Моделиране на взаимодействия на биомолекули", 18-19 февруари, Български суперкомпютърен център, 2012 София, БЪЛГАРИЯ;
- Training School "New Drugs for Neglected Diseases", 15-20 October, 2012 Siena, ITALY.

Заключение: Предложения за рецензия труд, притежава качествата на докторска дисертация и на автора Радослав Стефанов Мавревски следва да бъде присъдена научната степен „доктор" по: *4.6*

Информатика и компютърни науки.Докторска програма

“Информатика”

20.06.2015 подпис: 

София

Table 1. *Computational results obtained for 9 HP sequences*

Length	Sequences	Contacts	Time (sec.)
20	$HPHP_2H_2PHP_2HPH_2P_2HPH$	9	16.141
24	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	9	216.556
25	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	6	44.376
36	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	13	58.394
48	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	19	76.946
50	$H_2PHPHPHPH_4PHP_3HP_3HP_4HP_3HP_3HPH_4PHPHPHPH_2$	18	1.888
60	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	34	3.545
64	$H_{12}PHPHP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HPHPH_{12}$	34	17.336
102	$PH_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HPHP_2HP_5H_3P_4H_2PH_2P_5$ $H_2P_4H_4PHP_8H_5P_2HP_2$	28	7.438
123	$P_2H_3PHP_4HP_5H_2P_4H_2P_2H_2P_4HP_4HP_2HP_2H_2P_3H_2PHPH_3P_4H_3P_6H_2$ $P_2HP_2HPHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2PHPHPHPH$	34	431.562
136	$HP_5HP_4HPH_2PH_2P_4HPH_3P_4HPHPH_4P_{11}HP_2HP_3HPH_2P_3H_2P_2HP_2$ $HPHPHP_8HP_3H_6P_3H_2P_2H_3P_3H_2PH_5P_9HP_4HPHP_4$	36	530.455

Table 2 shows the results described in the literature, received during the performance of other algorithms on the first 8 HP sequences listed in Table 1. Also show the optimal number of contacts for these sequences [9,7,6].

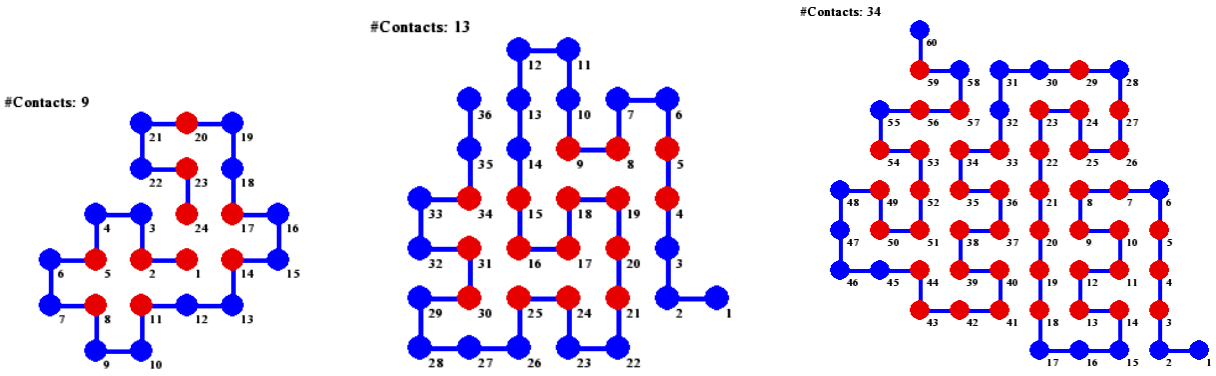
Table 2. *Comparison of the four algorithms*[a]

Length	Contacts				
	Optimal	Core	MC	GA	MS
20	9	9	9	9	9
24	9	9	9	9	9
25	8	6	7	8	8
36	14	13	12	14	14
48	23	19	18	22	22
50	21	18	19	21	21
60	36	34	31	34	34
64	42	34	31	37	38

[a]*Monte Carlo – MC* [7]*, Mixed Search – MS* [*7*]*, Genetic Algorithm – GA* [*7,6*]

As we can see from the table, the "*Core*" algorithm finds solutions that are very close to optimal. Even for some sequences we find the optimal solution. For long sequences with a length 102, 123 and 136 amino acids, the algorithm can find nearly optimal conformation. Note that the prediction of folds for long protein sequences is a difficult problem, since their natural state can only be obtained through several methods.

The resulting folds for sequences with length 24, 36 and 60 amino acids are given on Figure 3.



(*a*) (*b*) (*c*)

Figure 3: *Folding of proteins with lengths:* (*a*) *24 amino acids – 9 contacts,* (*b*) *36 amino acids – 13 contacts and* (*c*) *60 amino acids – 34 contacts.*

Conclusion

Computational experiments show that the algorithm is effective both for small and large scale protein sequences. When the entry sequence is large, usually we get solutions that are close to the optimal minimum energy conformation. Ahead of us stand the challenge to implement the algorithm on large-scale *HP* models, since in these cases the spatial structure grows rapidly with increasing chain length. Also, we can to improve the quality of folds, obtained from the proposed method by adapting to other lattice models or insertion of other techniques for analysis of protein structure.

With slight modification, this algorithm can be extended to 3D space. We note that coding of this algorithm is not complicated, and thus can be easily applied in practice.

REFERENCES

[1] B. Alberts, D Bray, A. Johnson et al., Essential Cell Biology: An Introduction to the Molecular Biology of the Cell, Garland Science Publishing, New York, 1998.

[2] B. Berger, T. Leighton, Protein folding in the hydrophobic-hydrophilic (hp) is np-complete, Proceedings of the second annual international conference on Computational molecular biology, New York, USA, (1998) 30-39.

[3] H. Yoon, Optimization Approaches to Protein Folding, PhD Thesis, School of Industrial and System Engineering, Georgia Institute of Technology, 2006.

[4] K. A. Dill, Theory for the folding and stability of globular proteins, Biochemistry-US 24 (1985) 1501-1509.

[5] K. A. Dill, S. Bromberg, K. Yue et al, Principles of protein folding – a perspective from simple exact models, Protein Sci. 4 (1995) 561-602.

[6] L. Toma, and S. Toma, Contact interactions method: A new algorithm for protein folding simulations, Protein Sci. 5 (1996) 147-153.

[7] M. Chen, and W. Huang, A. Branch and Bound Algorithm for the Protein Folding Problem in the HP Lattice Model, Genomics Proteomics Bioinformatics Vol. 3 No. 4. (2005) 225-230.
[8] M. Jiang, B. Zhu, Protein folding in the hexagonal lattice in the hp model, J. Bioinform. Comput. Biol. 3 (2005) 19-34.
[9] N. Ahn, S. Park, Finding an upper bound for the number of contacts in hydrophobic-hydrophilic protein structure prediction model, J. Comput. Biol. Vol. 17 No. 4 (2010) 647-656.
[10] R. Carr, W.E. Hart, and A. Newman, Discrete optimization models for protein folding, Technical report, Technical report, Sandia National Laboratories, 2003.
[11] S. Istrail, A. Hurd, R. Lippert, B. Walenz, S Batzoglou, J. H. Conway, F. W. Peyerl, Prediction of self-assembly of energetic tiles and dominoes: Experiments, mathematics, and software, Technical report, Sandia National Laboratories, 2000.
[12] S. Istrail, F. Lam, Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results, Communications in Information and Systems Journal 2009.
[13] V. Chandru, M.R. Rao, and G. Swaminathan, Protein folding on lattices: an integer programming approach, Society for Industrial Mathematics, Philadelphia, (2004) 185-196.
[14] Y. Duan, P. A. Kollman, Computational protein folding: From lattice to all-atom, IBM Syst. J. 40 (2001) 297-309.
[15] Z. Michalewicz, and D. B. Fogel, How to Solve It: Modern Heuristics, Springer, 2004.