Computational and Mathematical Approaches for Bioinformatics and Biophysics

生物信息和生物物理中的计算和数学方法

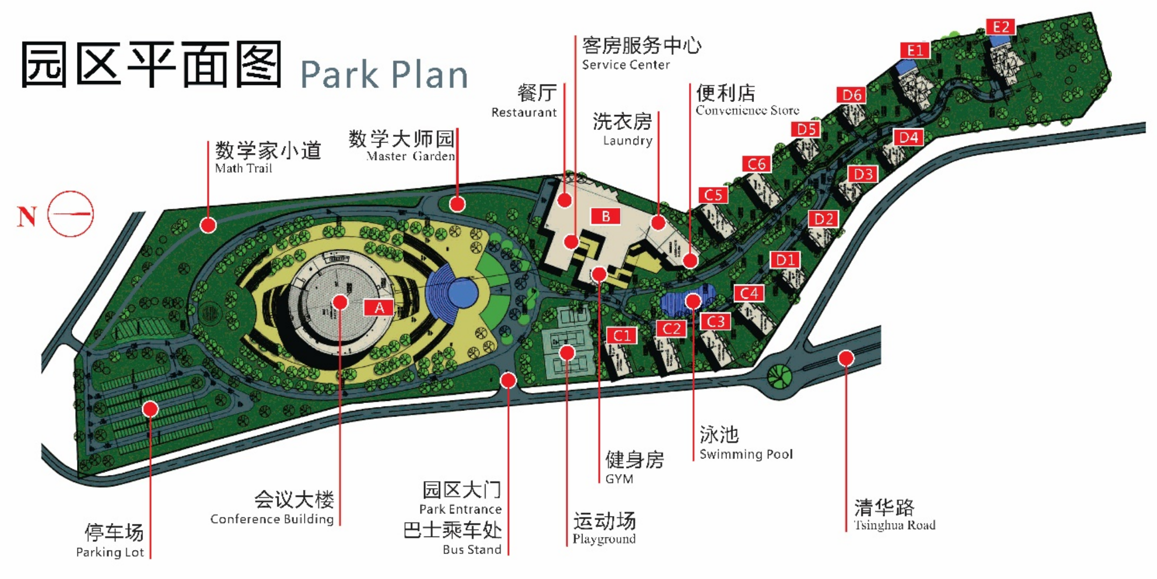
Dec. 10-14, 2018

Welcome to TSIMF

The facilities of TSIMF are built on a 23-acre land surrounded by pristine environment at Phoenix Hill of Phoenix Township. The total square footage of all the facilities is over 29,000 square meter that includes state-of-the-art conference facilities (over 10,000 square meter) to hold many international workshops simultaneously, two libraries, a guest house (over 10,000 square meter) and the associated catering facilities, a large swimming pool, workout gym and sport courts and other recreational facilities.

Mathematical Sciences Center (MSC) of Tsinghua University, assisted by TSIMF's International Advisory Committee and Scientific Committee, will take charge of the academic and administrative operation of TSIMF. The mission of TSIMF is to become a base for scientific innovations, and for nurturing of innovative human resource; through the interaction between leading mathematicians and core research groups in pure mathematics, applied mathematics, statistics, theoretical physics, applied physics, theoretical biology and other relating disciplines, TSIMF will provide a platform for exploring new directions, developing new methods, nurturing mathematical talents, and working to raise the level of mathematical research in China.

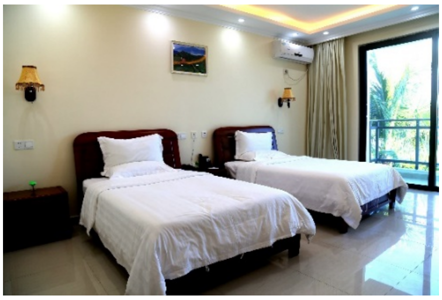
About Facilities



Registration

Conference booklets, room keys and name badges for all participants will be distributed at the front desk. Please take good care of your name badge. It is also your meal card and entrance ticket for all events.

Guest Room



All the rooms are equipped with: free Wi-Fi (no password), TV, air conditioner and other utilities.

Family rooms are also equipped with kitchen and refrigerator.

Library

Opening Hours: 09:00am-22:00pm

TSIMF library is available during the conference and can be accessed by using your room card. There is no need to sign out books but we ask that you kindly return any borrowed books to the book cart in library before your departure.

In order to give readers a better understanding of the contributions made by the Fields Medalists, the library of Tsinghua Sanya International Mathematics Forum (TSIMF) instituted the Special Collection of Fields Medalists as permanent collection of the library to serve the mathematical researchers and readers.

So far, there are 234 books from 47 authors in the Special Collection of Fields Medalists of TSIMF library. They are on display in room A220. The participants are welcome to visit.

Restaurant

All the meals are provided in the restaurant (Building B1) according to the time schedule.

Breakfast	07:30-08:30
Lunch	12:00-13:30
Dinner	17:30-19:00

Laundry

Opening Hours: 24 hours

The self-service laundry room is located in the Building 1 (B1).

Gym

The gym is located in the Building 1 (B1), opposite to the reception hall. The gym provides various fitness equipment, as well as pool tables, tennis tables etc.

Playground

Playground is located on the east of the central gate. There you can play basketball, tennis and badminton. Meanwhile, you can borrow table tennis, basketball, tennis balls and badminton at the reception desk.

Swimming Pool

Please note that there are no lifeguards. We will not be responsible for any accidents or injuries. In case of any injury or any other emergency, please call the reception hall at +86-898-38882828.

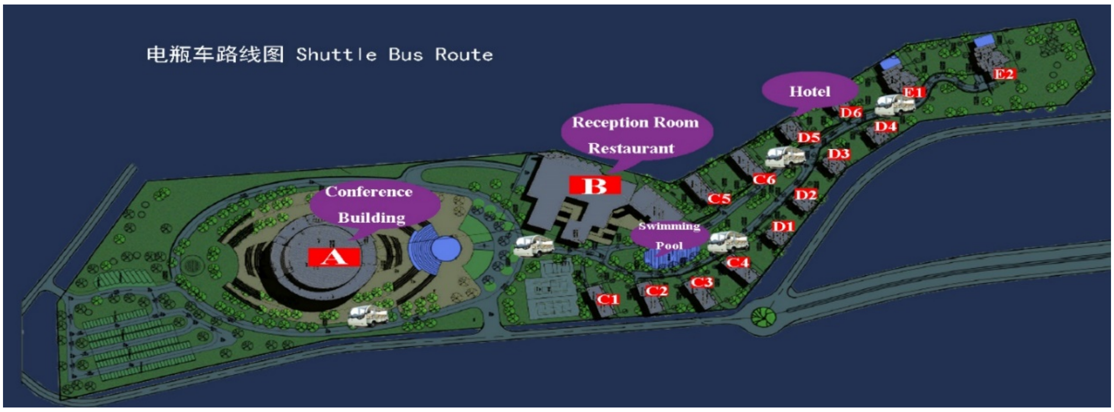
Outside Shuttle Service

We have shuttle bus to take participants to the airport for your departure service. Also, we would provide transportation at the Haihong Square （海虹广场） of Howard Johnson for the participants who will stay outside TSIMF. If you have any questions about transportation arrangement, please feel free to contact Ms. Li Ye (叶莉), her cell phone number is (0086)139-7679-8300.

Free Shuttle Bus Service at TSIMF

We provide free shuttle bus for participants and you are always welcome to take our shuttle bus, all you need to do is wave your hands to stop the bus.

Destinations: Conference Building, Reception Room, Restaurant, Swimming Pool, Hotel etc.



Contact Information of Administration Staff

Location of Conference Affair Office: ***Room 104, Building A***
Tel: 0086-898-38263896
Technical Supervisor: Mr.Shouxi,HE 何守喜
Tel: 0086-186-8980-2225
Email: hesx@ tsimf.cn

Conference Manager: Ms. Xianying, WU 吴显英
Tel:0086-186-8962-3393
Email: wuxianyingjojo@163.com

Location of Accommodation Affair Office: Room 200, Building B1
Tel：0086-898-38882828
Accommodation Manager: Ms. Li YE 叶莉
Tel: 0086-139-7679-8300
Email: yeli@tsimf.cn

Assistant Director of TSIMF
Kai CUI 崔凯
Tel/Wechat: 0086- 136-1120-7077
Email :cuikai@tsimf.cn

Director of TSIMF
Prof.Xuan GAO 高瑄
Email: gaoxuan@tsinghua.edu.cn

Schedule for Computational and Mathematical Approaches for Bioinformatics and biophysics Workshop, December 10-14, 2018

Time&Date	Monday (Dec. 10)	Tuesday (Dec. 11)	Wednesday(Dec. 12)	Thursday(Dec. 13)	Friday(Dec. 14)
7:30-8:30	*Breakfast (60 minutes)*				
Chair	Stephen Yau (5 minute open remarks)	John Zhang	Julie Mitchell	Chun Liu	
8:45-9:30	En-Bing Lin	Jie Liang	Shi Huang	Weihua Geng	Discussion
9:30-9:55	Yiming Bao	Minghui Yang	Jian Zhao	Xin Zhao	Discussion
9:55-10:25	*Group Photo about 5 minutes*	*Coffee Break*			
10:25-11:10	Changchuan Yin	Guowei Wei	Kelin Xia	Jia Wen	Discussion
11:10-11:35	Lei Zhang	Benzhou Lu	Jinzhi Lei	He Huang	Discussion
11:35-12:00	Feng Gao	Duan Chen	Yuyan Zhang	Rui Dong	Discussion
12:00-13:30	*Lunch (90 minutes)*				
Chair	Guowei Wei	Yingkai Zhang	Free Discussion 13:30-17:00 As for the sightseeing	Jie Liang	*Departure*
13:30-14:15	John Zhang	Chun Liu		Zhuomaji	
14:15-14:40	Yingkai Zhnag	Haipeng Gong		Rong He (Rui Dong)	
14:40-15:05	Zhijie Tan	Kun Tian		Yu Chen	
15:05-15:35	*Coffee Break (5 min Photo)*	*Coffee Break*		*Coffee Break*	
15:35-16:20	Julie Mitchell	Yi Xiao		Qi Wu	
16:20-16:45	Yongcheng Zhou	Steve Yau (1)		Lily He	
16:45-17:10	Discussion	Steve Yau (2)		Discussion	
17:00-17:30	Discussion	Steve Yau (3)		Discussion	
17:30	*Banquet 18:00-20:00*	Dinner			

Continuum modeling of selective ion permeation in a channel/nanopore

Benzhuo Lu
Institute of Computational Mathematics and Scientific/Engineering Computing,
Chinese Academy of Sciences, Beijing 100190, China.

Abstract

The Poisson-Nernst-Planck (PNP) model describing electrodiffusion processes can qualitatively capture some macroscopic properties of certain ion channel systems such as current-voltage characteristics, conductance rectification, and inverse membrane potential. With incorporation of other effects ignored in the mean field PNP theory such as size-effect, and in partcular the ion solvation effect, it is capable of simulating selective permeation in such as potassium channels. Potassium channels are much more permeable to potassium than sodium ions, although potassium ions are larger and both carry the same positive charge. It is known that the dehydration effect (closely related to ion size) is crucial to selective permeation in potassium channels. We incorporated Born solvation energy into the PNP model (BPNP) to account for ion hydration/dehydration effects when passing through the inhomogeneous dielectric channel environments. The model was applied to study a cylindrical nanopore and a realistic KcsA channel, and three-dimensional finite element simulations were performed. The BPNP model can distinguish different ion species by ion radius and predict selectivity for K+ over Na+ in KcsA channels. Furthermore, ion current rectification in the KcsA channel was observed by both the PNP and BPNP models. The I-V curve of the BPNP model for the KcsA channel indicated an inward rectifier effect for K+ (rectification ratio of 3/2) but indicated an outward rectifier effect for Na+ (rectification ratio of 1/6). These phenomena can be properly explained by the electrostatic energy landscape of the permeative ion along the channel resulted from the BPNP model. In addition, some other related models and applications will also be discussed in this talk.

INTERFACE METHODS FOR IMPLICITLY SOLVATED BIOMOLECULAR SIMULATION

WEIHUA GENG
DEPARTMENT OF MATHEMATICS
SOUTHERN METHODIST UNIVERSITY

Abstract: The Poisson-Boltzmann (PB) model is an effective implicit solvent approach for simulating solvated biomolecular systems. By treating the solvent with a mean field approximation and capturing the mobile ions with the Boltzmann distribution, the PB model largely reduces the degree of freedom and computational cost. However, solving the PB equation suffers from many numerical difficulties arising from interface jump conditions, complex geometry, charge singularities, and boundary conditions at infinity. In order to resolve these difficulties, we investigate interface methods under the frame of finite difference, boundary element, and finite element. We summarize our recent contributions in developing interface methods under various frameworks and provide insights for pros and cons of these methods. In addition, we provide several simulations for the calculations of important biological quantities such as solvation energy, binding energy, and pKa values using interface methods as well as a newly developed machine learning scheme.

Improving the conformational sampling for protein structural prediction

Tong Wang, Wenzhi Mao, Wenze Ding and Haipeng Gong*

School of Life Sciences, Tsinghua University, Beijing, China

Abstract

Machine learning techniques have been extensively used to facilitate the protein structure prediction nowadays. In this work, I will introduce our application of machine learning methods in improving the fragment selection and contact prediction, both of which will benefit the conformational sampling of protein structure prediction. In the first aspect, we constructed machine-learning models to optimize the extraction of near-native templates for fragments of 7-15 residues in the target protein. Fragment templates collected using our method show significant improvement in the degree of structural similarity to native ones over the other state-of-the-art methods and thus could enhance the efficiency of structure prediction algorithms using the fragment assembly protocol. In the second aspect, we developed a few machine-learning models either to predict the native residue contacts or to further refine the accuracy of a predicted residue contact map. Our methods show better or at least comparable performance to the other state-of-the-art ones. The predicted native residue contacts can be properly utilized in simulations to restrict the conformational space and thus to improve sampling efficiency in practical protein structure prediction.

*hgong@tsinghua.edu.cn

Mathematical modeling to stochastic gene expression, epigenetic regulation, and heterogenous stem cell regeneration

Jinzhi Lei
Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, China.

Abstract This talk will provide a general introduction of modeling the heterogeneity of cells in terms of stochastic gene expression and the random inheritance of epigenetic regulations. Epigenetic modifications, such as histone modification and DNA methylations, is essential for the regulation of gene expression, and can under go random changes over cell cycling. Changes in epigenetic modifications are important events in stem cell differentiation and plasticity. Dysregulations in the epigenetic modifications have been found association with many disease, including aging and cancer development. In this talk, I will present mathematical formulation to model stochastic gene expression that combine with epigenetic regulations and the random changes of epigenetic modifications over cell cycling, which lead to the heterogenous stem cell regeneration.

References

1. Jinzhi Lei, Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters. *Journal of Theoretical Biology*, **(**256)(2009), 485-492.

2. Romain Yvinec, Changjing Zhuge, Jinzhi Lei, Michael C. Mackey, Adiabatic reduction of a model of stochastic gene expression with jump Markov process. *Journal of Mathematical Biology*, **68**(2014), 1051-1070.

3. Xuan Zhang, Huiqin Jin, Zuoqin Yang, Jinzhi Lei, Effects of elongation delay in transcription dynamics. *Mathematical Biosciences and Engineering*, **11**(6)(2014), 1431-1448.

4. Wenjun Xia, Jinzhi Lei, Formulation of the protein synthesis rate with sequence information. *Mathematical Bioscience and Engineering*, **15**(2), 2018.

5. Rongsheng Huang, Jinzhi Lei, Dynamics of gene expression with positive feedback to histone modifications at bivalent domains. *International Journal of Modern Physics B*, **32**(2018), 1850075.

6. You Song, Honglei Ren, Jinzhi Lei, Collaboration between CpG sites in DNA methylation. *International Journal of Modern Physics B*, **31**(2017), 1750243.

7. Xiaopei Jiao, Jinzhi Lei, Dynamics of gene expression based on epigenetic modifications, *Communications in Information and Systems*, **18**(3)(2018), 125-148.

8. Rongsheng Huang, Jinzhi Lei, Cell-type switches induced by stochastic histone modification inheritance. *Discrete and Continuous Dynamical Systems-B* (to appear).

9. Jinzhi Lei, Simon A. Levin, Qing Nie, Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. *Proc. Natl. Acad. Sci. USA*, **111**(2014) E880-E887.

10. Qiaojun Situ, Jinzhi Lei, A mathematical model of stem cell regeneration with epigenetic state transitions. *Mathematical Bioscience and Engineering*, **14**(5&6)(2017), 1379-1397.

11. Jinzhi Lei, Qing Nie, Dong-bao Chen, A single-cell epigenetic model for paternal psychological stress-induced transgenerational reprogramming in offspring. *Biology of Reproduction*, **96**(6)(2018), 846-855.

12. Yucheng Guo, Qing Nie, Adam L. MacLean, Yanda Li, Jinzhi Lei, Shao Li, Multiscale modeling of inflammation-induced tumorigenesis reveals competing oncogenic and oncoprotective roles for inflammation. *Cancer Research*, **77**(22)(2017), 6429-6441.

Natural distinct inter-preference between genetic codon and protein secondary structure combinations

Zhuomaji, Xinqi Gong [*]

Abstract: The central dogma of molecular biology describes the process of genetic information transferred to protein. Many studies have found that genetic codon not only influences the protein amino acid sequence, but also affects protein 3D structure, such as local protein 3D structure may be affected by synonymous codon preferred usage. Here, in addition to the effect of single codons, we furtherly considering the preferences of short codon sequences for protein secondary structures. Also, we studied the preferences of short protein secondary structures for codon sequences. We studied in six cases that how codon combinations with length of N (N-codons) affect protein secondary structure element combinations with the same length (N-secondary structures), where $N = 1, \cdots, 6$. A few distinct codon combination sequences and their corresponding structure sequences were found by calculating Relative Codon Usage (RCU) and Relative Structure Usage (RSU). The preferences of many codon combinations vary for secondary structure combinations when N is different; similar preference patterns were found for protein secondary structure preference for genetic codons. This work is based on the CSandS database. In order to further confirm our conclusion, we selected seven proteins of human that are not in the CSandS to predict its secondary structures from nucleotide sequences using the statistical results. Prediction accuracy can reach 75.72%. It is sufficient to show the imprint of codons on protein structure, and it also indicates that codon usage probably is related to species.

Keywords: codon; short codon sequence; protein secondary structure; short structure sequence; distinct codon combination; predicting secondary structure;

[*]Institute for Mathematical Sciences, Renmin University of China, Beijing 100872, China. E-mail: xinqigong@ruc.edu.cn

Wavelet Analysis of Candidate Genes for Diabetic Retinopathy

En-Bing Lin
Department of Mathematics
Central Michigan University
Mt. Pleasant, MI 48859, USA

Abstract

Diabetic retinopathy, a major cause of adult blindness, is a medical condition in which damage occurs to the retina due to diabetes mellitus. Based on the Diaretinopathy database, we perform wavelet analysis on several candidate genes responsible for causing diabetic retinopathy. We obtain approximation and detail information of the numerical representations of cadidate genes. We compute discrete and continous wavelet transforms of each gene. We compare the computational and graphical results. We also perform wavelet analysis and wavelet transform on the Fibroblast Growth Factor 21 gene which can be included in gene therapy. Through this study, it is anticipated to provide better disease managements and ensure better prognosis in diabetic retinopathy.

A multiscale virtual particle based elastic network model for biomolecular normal mode analysis

Kelin Xia

In this talk, I will discuss our works on multiscale virtual particle based elastic network models (MVP-ENMs), which includes MVP based Gaussian network model (MVP-GNM) and MVP based anisotropic network model (MVP-ANM). The multiscale virtual particle (MVP) model is proposed for the discretization of biomolecular density data. With this model, large-sized biomolecular structures can be coarse-grained into virtual particles such that a balance between model accuracy and computational cost can be achieved. Further, a particle-and-distance dependent spring parameter is used in MVP-ENMs to better characterize the interactions within the model. The MVP-GNM has been tested in the prediction of Debye-Waller factors (B-factors) with comparably good accuracy as GNM. MVP-ANM can deliver highly accurate low-frequency eigenmodes for Cryo-EM data. Kelin Xia (NTU) Sanya.

Moreover, complex multiscale virtual particle (CMVP) model is designed to characterize the intrinsic differences between the various components in biomolecular complexes. CMVP-GNM has been proposed to improve the accuacry of GNM in B-factor predictions of protein-nucleic acid complexes. CMVP-ANM is used to model collective motions of the virus structures that are composed of protein capsid and genome region.

A new framework for studying genetic diversity, phylogenetics, and complex traits

Shi Huang, Center for Medical Genetics, Central South University, Changsha, Hunan, China

Our maximum genetic diversity (MGD) hypothesis was inspired by one of the most astonishing findings in modern science, the genetic equidistance phenomenon first discovered in 1963, that has been originally mis-interpreted by the molecular clock and in turn the modern evolutionary theory as authored by Kimura and Darwin. The hypothesis posits that evolution involves two different if not opposite processes, micro- and macro-evolution. Each species has a specific level of complexity as defined by the number of cell types and a corresponding level of maximum tolerable level of genetic diversity or random errors in the genome. There exists a self-evident inverse relationship between maximum genetic diversity tolerable to a species and the phenotypic complexity of the species, which is equivalent to the intuition that the more complex/ordered the system the more precise the building blocks.

Micro-evolution is minor and slow diversification of species involving random mutations within allowed ranges of genetic diversity plus genetic drift or natural selection as described by Kimura and Darwin. Macro-evolution is major and rapid increase in phenotypic or epigenetic complexity with a corresponding decrease or suppression in the maximum allowed level of genetic diversity or range of errors permissible in the genome of the newly evolved species. The MGD theory considers the genome to be largely devoid of neutral sites, which has solved the century old genetic diversity riddle and been useful in studying complex traits and diseases and rewriting phylogenetics.

References:

Hu, T., Long , M., Yuan D., Zhu Z., Huang, Y., and Huang, S. (2013) The genetic equidistance result: misreading by the molecular clock and neutral theory and reinterpretation nearly half of a century later. Sci China Life Sci, 56：254-261.

He, P., Lei, X., Yuan, D., Zhu, Z., and Huang, S. (2017) Accumulation of minor alleles and risk prediction in schizophrenia. Sci. Rep. doi:10.1038/s41598-017-12104-0

Huang, S. (2012) Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci, 55: 709-725.

Huang, S. (2016) New thoughts on an old riddle: what determines genetic diversity within and between species. Genomics, 108: 3-10. doi:10.1016/j.ygeno.2016.01.008

Lei, X., Yuan, D., and Huang, S. (2018) Collective effects of common SNPs and risk prediction in lung cancer. Heredity, doi:10.1038/s41437-018-0063-4

Yuan, D., Lei, X., Gui, Y., Zhu, Z., Wang, D. Yu, J., and Huang, S. (2017) Modern human origins: multiregional evolution of autosomes and East Asia origin of Y and mtDNA. bioRxiv. doi: https://doi.org/10.1101/101410

Constructing contact maps of protein and RNA structures using direct coupling analysis

Yi Xiao

School of Physics, Huazhong University of Science and Technology, Wuhan 430074, China

During evolution, the residues in direct contact in protein and RNA tertiary structures are more likely correlated through co-evolution in order to maintain their structures and functions. These coevolutionary pairwise residue couplings have been used to identify binding sites and predict tertiary structures of protein and RNA.

Accuracies of contact predictions strongly depend on the used models. For examples, mutual information (MI) of a multiple sequence alignment (MSA) was used as a measurement of pair correlations. The shortage of this measure is that the predicted contacts contain many pairwise residues that are not in direct contacts in tertiary structure, resulting in many false positives. To solve this problem, direct coupling analysis (DCA) has been proposed to disentangle direct contacts from indirect ones. There are different versions of DCA that use different approximations and algorithms. However, these DCA algorithms still give many false positives and need further improvements. Here we give a detailed analysis of the performance of some DCA algorithms in contact inference and propose some ways of picking out more true positives.

References

1. Morcos F, Hwa T, Onuchic JN, Weigt M: **Direct coupling analysis for protein contact prediction**. *Methods Mol Biol* 2014, **1137**:55-70.
2. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation**. *Nat Biotechnol* 2012, **30**(11):1072-1080.
3. Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y, Xiao Y: **Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis**. *Nucleic Acids Res* 2017, **45**(11):6299-6309.
4. Huang Y, Li H, Xiao Y: **3dRPC: a web server for 3D RNA-protein structure prediction**. *Bioinformatics* 2018, **34**(7):1238-1240.

Mathematical deep learning for drug discovery

Guowei Wei
Department of Mathematics
Michigan State University, East Lansing, MI 48824, USA

Designing efficient drugs for curing diseases is of essential importance for the 21st century's life science. Computer-aided drug design and discovery has obtained a significant recognition recently. However, the geometric complexity of protein-drug complexes remains a grand challenge to conventional computational methods, including machine learning algorithms. We assume that the physics of interest of protein-drug complexes lies on low-dimensional manifolds or subspaces embedded in a high-dimensional data space. We devise topological abstraction, differential geometry reduction, graph simplification, and multiscale modeling to construct low-dimensional representations of biomolecules in massive and diverse datasets. These representations are integrated with various deep learning algorithms for the predictions of protein-ligand binding affinity, drug toxicity, drug solubility, drug partition coefficient and mutation induced protein stability change, and for the discrimination of active ligands from decoys. I will briefly discuss the working principle of various techniques and their performance in D3R Grand Challenges, a worldwide competition series in computer-aided drug design and discovery (http://users.math.msu.edu/users/wei/D3R_GC3.pdf).

A Biomolecular Case Study in Feature Selection for Machine Le arning

Julie Mitchell

A common question asked of machine learning algorithms is which of its features/descriptors is “best.” Here, we examine feature selection using features from the KFC2 model for protein-protein interface hot spots and features derived from the Rosetta molecular modeling package. Intercorrelations among features reflect the underlying physics but complicate the question of feature importance. By studying feature selection with multiple strategies, important principles emerge even in the absence of a clear feature ranking.

And Biography:

Julie Mitchell is the Deputy Director of Biosciences and Chief
Scientist for Computational Biology at Oak Ridge National Laboratory. She was previously a Professor of Mathematics and Biochemistry at the University of Wisconsin – Madison. Mitchell grew up in the San Francisco Bay Area and did her PhD in Mathematics at UC Berkeley, later migrating to the computational/math biology field as a postdoc at UC San Diego.

The Z-curve theory for the graphical representation of DNA sequences and its application in genome analysis

Feng Gao

Integrated molecular modeling and machine learning to target protein-protein interactions

Yingkai Zhang

Protein-protein interaction (PPI) interfaces generally feature large, flat and dynamic binding surfaces, which pose a challenge for conventional computational analysis tools. In this talk, I will describe a topographical mapping approach to reveal a fragment-centric modularity at PPI interfaces. The resulting high-resolution map of underutilized, targetable pocket space can be used to guide the rational design and optimization of novel inhibitors. In addition, I will describe our recent progresses to improve docking scoring functions with machine learning.

Free energies in protein-protein and protein-ligand bindings

John Z.H. Zhang
East China Normal University & NYU Shanghai
John.zhang@nyu.edu

Theoretical calculation of protein-protein and protein-ligand binding free energies is a grand challenge in computational biology. Accurate prediction of critical residues along with their specific and quantitative contributions to protein-protein binding free energy is extremely helpful to reveal binding mechanisms and identify drug-like molecules that alter protein-protein interactions. In this talk we develop an efficient approach (Interaction Entropy) to computing quantitative residue-specific contributions to protein-protein and protein-ligand binding free energies. The approach provides explicit contribution of the entropic loss in binding free energy of individual residues directly from fluctuation of the interaction energy in MD simulation. Studies for an extensive set of realistic protein-protein interaction systems and for specific protein-ligand binding systems showed that by including the entropic contribution, the computed residue-specific binding free energies are in better agreement with the corresponding experimental data. Predictions of hot stops for some important protein-protein interactions are discussed.

References

1. Duan, L.L., X. Liu, and J.Z.H. Zhang, *Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein-Ligand Binding Free Energy*. J. Am. Chem. Soc., 2016. **138**(17): p. 5722-5728.
2. Sun, Z.; Yan, Y.; Yang, M.; Zhang, J.Z.H., *Interaction entropy for protein-protein binding*. J. Chem. Phys., 2017. **146,** 124124.
3. Yan, Y.; Yang, M.; Ji, C.; Zhang, J.Z.H., *Interaction Entropy for Computational Alanine Scanning*. J. Chem. Inf. Model., 2017. **57**, 1112–1122.
4. Liu, X.; Peng, L.; Zhou, Y.; Zhang, Y.; Zhang, J.Z.H.; *Computational Alanine Scanning with Interaction Entropy for Protein–Ligand Binding Free Energies,* J. Chem. Theo. Comput., 2018, **14** (3), 1772–1780.
5. Song J, Qiu L, Zhang JZH. *An efficient method for computing excess free energy of liquid.* Sci China Chem, 2018, 61: 135–140.

Single nucleotide polymorphisms genotyping of \textit{S. aureus}] {Whole genome single nucleotide polymorphisms genotyping of \textit{Staphylococcus aureus}}

Changchuan Yin, Stephen S.-T. Yau

Next-generation sequencing technology enables routine detection of bacterial pathogens for clinical diagnostics and genetic research. Whole genome sequencing has been of importance in the epidemiologic analysis of bacterial pathogens. However, few whole genome sequencing-based genotyping pipelines are available for practical applications. Here, we present the whole genome sequencing-based single nucleotide polymorphisms (SNPs) genotyping method and apply to the evolutionary analysis of methicillin-resistant \textit{Staphylococcus aureus}. The SNP genotyping method calls genome variants using next-generation sequencing reads of whole genomes and calculates the pair-wise Jaccard distances of the genome variants. The method may reveal the high-resolution whole genome SNP profiles and the structural variants of different isolates of methicillin-resistant \textit{S. aureus} (MRSA) and methicillin-susceptible \textit{S. aureus} (MSSA) strains. The phylogenetic analysis of whole genomes and particular regions may monitor and track the evolution and the transmission dynamic of bacterial pathogens.

Assessment of kmer degeneration method for complicated genomes

Shuai Liu[b,c,**], Shaojun Pei[a,**], Stephen S.-T. Yau[a,*], Qi Wu[c,*]

[a] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China*
[b] *Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, China*
[c] *Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China*

Abstract

The kmer frequency is widely used in alignment-free sequence analysis methods. To better describe the overall statistical features of a complicated sequence, such as those of mammals, a longer length of kmer is required. However, the long length of kmer will cause exponential increasing of the types of kmer (4^K types of kmer with length K), which results in an extremely intensive computational burden and makes k-mer method impractical. In this work, we propose a novel method of kmer degeneration (KD) to balance the kmer length and kmer type. The method only considers N positions of nucleotides out of K positions of K-mers and degenerates all other $(K - N)$ positions. Then the K-mers can be substituted by $(K)N$-mer. Therefore, the kmer types were reduced from 4^K to 4^N and remain the linkages among the nucleotides within the K-mer. We first show how N can be determined for a given K. Then we assess which types of combinations of the N positions from the K positions are better for describing the sequence. Finally, to illustrate the utility of the method, we construct the phylogenetics tree of *Carnivora* with 16 genomes using our method, which is better than non-degenerated kmer with the same N value.

Keywords: kmer frequency, kmer degeneration, combination, mammalian

[*] co-corresponding author
[**] Shaojun Pei and Shuai Liu contributed equally
Email addresses: `yau@uic.edu` (Stephen S.-T. Yau), `ribozyme@ioz.ac.cn` (Qi Wu)

Preprint submitted to Journal of LaTeX Templates *December 2, 2018* genome

1. Introduction

Alignment-free (AF) sequence analysis methods focus on subsequences (words or kmers) of defined or varied lengths of the sequence [1–3]. The types[4], frequencies [5], or positions [6] of the kmers in the sequence are considered as characteristics to do genome comparison. Of the possible characteristics, frequency is the most widely used in alignment-free methods [7, 5, 2]. The algorithm was shown with the CVTree program as an example [5]. Starting with a given DNA sequence of length L, a sliding window of length K is run through the position 1 to $L - K + 1$ to count the frequency of subsequences (kmers or strings). The total possible types of such strings could be 4^K for DNA sequences. Denote the frequency of appearance of kmer $\alpha_1\alpha_2...\alpha_K$ by $f(\alpha_1\alpha_2...\alpha_K)$, where $\alpha_i \in \{A, C, G, T\}$. This frequency divided by the total number $L - K + 1$ ($K << L$) of K-mers in the given DNA sequence may be taken as the probability $p(\alpha_1\alpha_2...\alpha_K)$ of appearance of the kmer $\alpha_1\alpha_2...\alpha_K$ in the sequence:

$$p(\alpha_1\alpha_2...\alpha_K) = \frac{f(\alpha_1\alpha_2...\alpha_K)}{K - L + 1} \tag{1}$$

The collection of such frequencies or probabilities describes the overall statistical features of the sequence concerned, which could then be used for further analysis such as whole-genome phylogeny reconstruction.

Clearly, in this approach, the most important parameter is the length of kmer. Therefore, the length of kmer, or K value, should be determined by some types of prior information to obtain optimal target sequence resolution. One straightforward limit is that the K value must be far less than the sequence length, or L value [5]. Sims et al. (2009) derived the upper limits of kmer length by cumulative relative entropy (CRE) for a given symbol sequence, which represented the accuracy of predicting kmer frequencies for all lengths of kmer from Equation 1 [7]. They found that the length reaches the upper limit for use in genome comparison when the CRE approaches zero. In a qualitative view, the kmer length is related to both the sequence length and the kmer type number. With the long k-mer lengths, the kmer type number increases exponentially, which causes two problems. One problem is that a large number of kmer types will exceed the storage capacity of the computer. Another is that, if the number of kmer types far exceeds the amount of kmers appearing in the sequence, the majority of kmers will appear only once in non-repetitive sequences. This means, for any non-repetitive sequence, long kmer lengths result in a uniform distribution of the kmer frequencies.

In this work, we proposed k-mer degeneration (KD) method to balance the longer kmer length and the increased kmer types. In k-mer degeneration method, we only keep N ($N << K$) positions in the K positions of a sliding window, so that a K-mer is degenerated to an N-mer. Therefore, the kmer types are reduced from 4^K to 4^N. At the same time, the N positions of the $(K)N$-mer could have a span ranging of K. Therefore, the linkages among nucleotides within the K-mer is, to some extent, reflected through the degenerated $(K)N$-mer, so the information in the $(K)N$-mer could be greater than that in a naive N-mer. Using Shannon information [8] as a measurement, we assessed our method on the data set of mammalian genomes. First, for a given K, we used distribution of Shannon information to determine which N is suitable for $(K)N$-mer. Then we accessed different types of combinations of the N positions from the K positions and the results indicated that the cases of continuous N positions possess significantly less information than the cases of dispersed N positions for a fixed K. Finally, we constructed the phylogenetics tree of *Carnivora* with 16 genomes by using our method in the form of dispersed $N(K)$-mer, which is better than those using non-degenerated kmer with the same N value.

2. Materials and Methods

2.1. Genome data collection

Genome sequences were downloaded from public database. The selected species included human (*Homo sapiens*) and mouse (*Mus muscullus*) from the superorder of *Euarchontoglires*, dog (*Canis lupus familiaris*) from the superorder of *Laurasiatheria*, elephant (*Loxodonta africana*) from the superorder of *Afrotheria*, and wallaby (*Macropus eugenii*) from the order of *Marsupialia*. Except for the wallaby, which is located at the basal branch of the mammal phylogenetic tree, the other 4 species belong to the placentals, which is at the crown of the mammalian tree. The sequence length was fixed to 50 Kbps in this work. For different genomes, 10 segments of genome sequence were independently sampled from the genome.

To construct phylogenetic tree and evaluate the performance, we picked the mouse and human genomes to make up an outgroup, and red panda(*Aliurus fulgen*), ferret(*Mustela putorius furo*), giant panda (*Ailuropoda melanoleuca*), tiger(*Panthera tigris*), polar bear(*Ursus maritimus*), cat(*Felis catus*), pacific walrus(*Odobenus rosmarus*), sea otter (*Enhydra lutris*), brown bear(*Ursus arctos*), weddell seal(*Leptonychotes weddellii*), cheetah(*Acinonyx jubatus*), puma(*Puma concolor*), leopard (*Panthera pardus*) from *Carnivora* (Table S3).

2.2. kmer degeneration (KD) Method

For a DNA sequence with length L, a window with length K is run through the sequence from the beginning with steps of one nucleotide. One can obtain $L-K+1$ kmers appearing in the sequence, which belong to T_K types of kmers. It is always the case that T_K is smaller than $L-K+1$, since there must be some kmers appearing more than once in the sequence. At the same time, the total number of kmer types would be 4^K , giving

$$\begin{cases} T_K \leq L-K+1 \\ T_K \leq 4^K \end{cases} \tag{2}$$

In k-mer degeneration method, we only keep the N positions of nucleotide in K positions of the sliding window. Except for the N positions in the window, all other positions in the window are degenerated. A new $(K)N$-mer was then obtained, where the term “$(K)N$-mer“ denotes the new kmer with length N that is degenerated other $(N-K)$ positions from the original K-mer (Figure 1). From a mathematical viewpoint, there are $\binom{K}{N}$ combinations to pick N positions out of K positions. When dealing with one sequence, one can fix one combination and perform kmer degeneration (KD) method for a sequence. The symbol T_N is used to denote the number of $(K)N$-mer types in this case. Notably, although the number of kmers appearing in the sequence is the same for both K-mer and $(K)N$-mer, the total number of kmer types is decreased from 4^K to 4^N. Therefore, one has

$$\begin{cases} T_K \leq L-K+1 \\ T_K \leq 4^K \\ T_N \leq 4^N \\ T_N \leq T_K \end{cases} \tag{3}$$

For a genome sequence segment with a length of 50 Kbp, one could analyze it with a kmer whose length is 50 bp. Consider the case that the 50-mer is degenerated into an 8 bp length of (50)8-mer, one has

$$\begin{cases} T_K \leq L-K+1 = 49951 \\ T_K \leq 4^K = 4^{50} \approx 1.28 \times 10^{30} \\ T_N \leq 4^N = 4^8 = 65536 \\ T_N \leq T_K \end{cases} \tag{4}$$

Therefore, one can see the difference between using K-mers directly and instead using degenerated $(K)N$-mers to analyze sequence. The values of 49951 and 65536 are the same order of magnitude, but 1.28×10^{30} is much larger. Supposed 49951 balls were placed into 1.28×10^{30} boxes, each ball has an equal probability to be placed into any of the boxes. In most cases, one may obtain a distribution much closer to a uniform distribution, since it would be very rare to place two or more balls into one box. By contrast, if one put 49951 balls into

65536 boxes, one could easily obtain a recognizable distribution that would be significantly different from a uniform distribution(Figure 2).

2.3. Definition of the dispersed and continuous $(K)N$-mers

In kmer method, for a DNA sequence of length L, there is a sliding window of length K which is run through the sequence. Our kmer degeneration method only picks up N positions out of the K positions of the sliding window. To remain the span of K-mer, the first and last position of the window should be picked up. Then one can pick up other $N-2$ positions randomly from remaining positions in the window. So there are $\binom{K-2}{N-2}$ combinations totally. In this study, we mainly studied two special cases. Two forms of $(K)N$-mers were defined the dispersed form and continuous form. For any form of $(K)N$-mer, let D_i be the length of the interval between every two nucleotides, where $i = 1, 2, ..., N-1$. Then Let

$$d = \left\lfloor \frac{K-N}{N-1} \right\rfloor \tag{5}$$

Definition 1. *If all the $d-1 < D_i < d+1,\ \ i = 1, 2, ..., N-1,$ we call these $(K)N$-mers as the dispersed form.*

For example, when $K = 7$ and $N = 3$, the d is $\frac{7-3}{3-1} = 2$; therefore, the three positions chosen for the $(K)N$-mer should be 1-3-7, 1-4-7, 1-5-7 (Fig. 3A). When $K = 8$ and $N = 3$, the d is also 2, the forms were: 1-3-8, 1-4-8, 1-5-8. With this method, we could produce a number of $(K)N$-mers as a group whose position was dispersed from the original K-mer.

Definition 2. *If all the centering $N-2$ positions are continuous, we call these $(K)N$-mers as the continuous form.*

For example, when $K = 7$ and $N = 4$, the continuous forms are 1-2-3-7, 1-3-4-7, 1-4-5-7, 1-5-6-7 (Fig. 3B).

2.4. Measurement for degenerated kmer method

As has been mentioned above, there are $\binom{K}{N}$ combinations for choosing N positions out of K positions. The question is whether there are any differences between these combination types. This question could be asked in the view of information theory: Which type of combination(s) could provide more information? The Shannon information is a well-defined index in this case. For a finite discrete random variable X with distribution $f(X) = p(x_i)$, where $i = 1, 2, ..., n$, the Shannon information $I(X)$ is

$$I(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{6}$$

When given a target sequence and the $(N)K$-mer, the number of $(N)K$-mer types T_N is determined. And the distribution is determined by all frequencies of the T_N types of kmer, so

$$I(N) = -\sum_{i=1}^{T_N} p_i \log_2 p_i \tag{7}$$

2.5. Generation of posterior distribution

The most straightforward measurement method would be to compute all of the combinations and compare them. However, for large values of K and N, such as the previous case of $K = 50$ and $N = 8$, the number of $\binom{K}{N}$ would be too large. We randomly sampled 1000 combinations from all combinations. For each type of combination of $(K)N$-mer, the frequencies of $(K)N$-mer were computed and then the Shannon information of this combination was calculated by the frequencies. The range from the maximum information value to the minimum information value was divided into 100 groups. The 1000 information values were then grouped to produce a posterior distribution of the information. With this posterior distribution, the differences among the combinations can be assessed.

2.6. Construction of phylogenetic trees

Phylogenetic tree is required to be constructed to test and verify the feasibility of kmer degeneration method. With a specific way of degenerating kmer, we could count and calculate the frequency of $K(N)$-mers. To eliminate the interference of random background, The frequency of random background should be subtracted by the frequency of K(N)-mers [9].

2.6.1. Frequency or Probability of Appearance of $(K)N$-mer

For a genome sequence of length L, we denote the frequency of appearance of the $(K)N$-mer $\alpha_1\alpha_2...\alpha_N$ by $f(\alpha_1\alpha_2...\alpha_N)$, where each $\alpha_i \in \{A, C, G, T\}$. This frequency divided by the total number $(L - K + 1)$ may be taken as the probability $p(\alpha_1\alpha_2...\alpha_N)$:

$$p(\alpha_1\alpha_2...\alpha_N) = \frac{f(\alpha_1\alpha_2...\alpha_N)}{L - K + 1} \tag{8}$$

2.6.2. Subtraction of Random Background

According to Markov model prediction, we subtract the random background from the sequence as follows.

Suppose we have done $(K)N - 1$-mer and $(K)N - 2$-mer The probability of $(K)N$-mer is predicted by:

$$p^0(\alpha_1\alpha_2...\alpha_N) = \frac{p(\alpha_1\alpha_2...\alpha_{N-1})p(\alpha_2\alpha_3...\alpha_N)}{p(\alpha_2\alpha_3...\alpha_{N-1})} \tag{9}$$

In order to make sure all the frequencies of different kmer lengths (N, $N - 1$ and $N - 2$) in Equation 9 are non-zero, we should scan the genomes checking the different N values and determine the maximal N which makes all kinds of kmer appear at least once in every genome. For the data set in this study, we can get the maximal $N = 11$ (Table S4).

2.6.3. (K)N-mer vector and distance metric

We define

$$a(\alpha_1\alpha_2...\alpha_N) = \frac{p(\alpha_1\alpha_2...\alpha_N) - p^0(\alpha_1\alpha_2...\alpha_N)}{p^0(\alpha_1\alpha_2...\alpha_N)}. \tag{10}$$

Then we can obtain the $(K)N$-mer vector for species A:

$$A = (a_1, a_2, ..., a_{4^N})$$

Likewise, we can get the $(K)N$-mer vector:

$$B = (b_1, b_2, ..., b_{4^N})$$

So the distance $D(A, B)$ between the two species is defined as

$$D(A, B) = \sqrt{\sum_{i=1}^{4^N}(a_i - b_i)^2} \tag{11}$$

2.6.4. Tree construction and decoration

MEGA, a popular software in molecular evolutionary genetic analysis, which could be employed to transform the distance matrix of species into phylogenetic tree by Neighbor Joining method. FigTree, which is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures, could help us polish up the phylogenetic tree.

3. Results

3.1. The best N of $(K)N$-mer for a given K

As mentioned in the methods above, it is necessary to determine the best N value for a fixed K, since an N value that is too small will give a very noisy distribution, while too large will be very close to the K value and weaken the effectiveness of KD method. To assess the performance of different N values, we introduced a posterior distribution by picking 1000 combinations from $\binom{K}{N}$ and constructed a histogram of the Shannon information of 1000 combinations. We randomly selected a 50 Kbp genome sequence ($L = 50000$) and used $K = 50$ as an example to examine different N values. The use of this L length followed the work of Gentles et al. (2001)[10]. The results showed that for small N values, the distribution had multiple peaks. Since $N = 8$, the distribution became a single-peak distribution (Figure 4). We re-examined this for 10 different 50 Kbp sequences from human genomes and proved the robustness of the results. Therefore, the result of $N = 8$ was determined as a reference length of $(K)N$-mer.

3.2. Dispersed $(K)N$-mers have more information than continuous $(K)N$-mers

We focused on dispersed and continuous forms of combinations in $\binom{K}{N}$, which are defined in Section 2.3. We randomly selected 10 segment sequences from the human genome, as above, to demonstrate the result with the parameters of $L = 50000$(bp), $K = 50$(bp) and $N = 8$(bp). The results are shown in Table 1. It can be seen that all continuous $(K)N$-mers have Shannon information that is distributed extremely to the left side of the distribution. The majority is smaller than 5% quantile. By contrast, the dispersed $(K)N$-mers are distributed to the right side of the distribution and mostly larger than 95% quantile. Further, to compare the two forms' Shannon information on genome sequence and simulated random sequence, we generated a nucleotide sequence of 50kb and selected a 50kb segment sequence of human genome. Then we enumerated all forms of (20)5-mer and calculated their Shannon information. We found Shannon information of dispersed and continuous forms on simulated sequence was almost the same. However, for genome sequence, their information was significantly different. According to Table 2, we also found Shannon information of dispersed forms on genome sequence was significantly bigger than that of continuous forms.

3.3. Influence of K value for $(K)N$-mer in humans

It is necessary to assess the influence of various K values for a given N between dispersed and continuous $(K)N$-mers. We performed the assessment with 10 different genome sequence segments as repeats with $L = 50000$ and $N = 8$. For each sequence we considered 6 types of K values above $K = 50$. Table 3 indicates the result of one segment of sequence in the human genome. The results of all the 10 repeats of segment sequences are presented in Table S1. The results were similar to those in Table 1, which indicated that the difference between the dispersed and continuous form was not markedly influenced by K values less than 200. For K values greater than 200, the dispersed form of the $(K)N$-mer appeared more in the 50-75% interval. When the K values approached the level greater than 300, the dispersed form of the $(K)N - mer$ began to appear in the other side of the peak, in the interval of 25-50%.

One might be curious whether there is a definite upper limit of information volume for kmers. In fact, when the total kmer types were far greater than sequence length, all of the frequencies could be much closer to $\frac{1}{L}$; therefore, the upper limit of information for a given sequence length would be $\log_2 L$. Notably, this value is independent from the K value. For the case of $L = 50000$ in this work, the upper limit of information I_{max} is

$$I_{max} = \log_2 L = \log_2 50000 = 15.6096 \tag{12}$$

The $(K)N$-mers were compared with $N = 8$ and various K values (Figure 5). It could be seen that the information increased slightly for kmers with lengths less than 200. For kmers longer than 200, the information tended to fluctuate. It was noticeable that the difference of Shannon information is significant in different genome segment sequences, which can reflect heterogeneity of the genome sequences.

3.4. Comparison of different $(K)N$-mer combinations in pan-mammals

We examined the difference between dispersed and continuous $(K)N$-mers with different mammalian genomes. Species were selected from the main branches of mammals. It can be seen from Table 4 (one segment of sequence from one species) and Table S2 (for all other sequences) that for all genome sequences from the 5 mammalian species, the information values from continuous combinations gathered far away to the left side of the distribution peak, while the values from dispersed combinations were located in the right side of the distribution peak.

3.5. Assessment of kmer degeneration method in tree construction

To test the availability of KD method in alignment-free phylogenomics approach, we constructed the phylogenic relation of *Carnivora* with 16 genomes as a case. The data set included 14 carnivorous species as well as human and mouse genome as outgroup. The N value was determined as 11 and we checked a series of K values of 50, 60, 70, 80, 90, 100, 110, 120, 130, 200, 250 and 300. The results showed that when K =50, 60, 70, 80, 90, 110, the topology of Canine suborder tree is the same with the general accepted phylogeny (Fig. 6A) [11]. What should be noticed is the position of the *Pinnipedia*. In some studies, it is sister group of the Weasel superfamily, while in another studies, it is sister group of it is sister group of *Arctoidea* superfamily [12, 13]. Our result suggests that it is sister group of the Weasel superfamily instead of the *Arctoidea* superfamily [11]. There are two kinds of topology of tree for the *Feliformia* suborder in our result. When K =50, 70, 80, 90 and 110, cheetah is clustered with the linage of *Panthera* genous while domestic cat and puma are clustered as one lineage. While in the case of K = 100, none of the three species, cheetah, puma and domestic cat, have been clustered (Fig. 6B). Compared with our alignment-free result, it has been generally accepted that the three species should be clustered into one monophyletic group (Fig. 6D) [14, 11]. Despite of this discrepancy, our result is better than those using non-degenerated kmer with the same N value (Fig. 6C).

4. Conclusions

The aim of this study was to enable the alignment-free whole-genome phylogeny for complicated large genomes, such as in mammals or birds. For the application of alignment-free methods, the large genome size confined the usage of long K-mers which have a vast amount of kmer types. Our KD method reduced the K-mer types while maintaining a long span as K-mer. This makes it possible to investigate whole-genome phylogeny with frequencies of long K-mers. For a given K, we first use the distribution of Shannon information to determine the best N for $(K)N$-mer. Then by analysis of different combinations of positions, we find that the cases of continuous N positions possess significantly less information than the cases of dispersed N positions for a fixed K. So for a DNA sequence, one can use dispersed $(K)N$-mer to substitute K-mer. We applied our method to Carnivora with 16 genomes, which is better than non-degenerated kmer with the same N value. Therefore, our method can be used to other genomes for future research.

5. Conflict of interest statement

The authors declare no competing financial interests

6. Acknowledgements

This work is supported by National Natural Science Foundation of China grant (#91746119).

7. References

[1] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment., Proceedings of the National Academy of Sciences of the United States of America 83 (14) (1986) 5155–5159.

[2] C. X. Chan, G. Bernard, O. Poirion, J. M. Hogan, M. A. Ragan, Inferring phylogenies of evolving sequences without multiple sequence alignment, Sci Rep 4 (39) (2014) 6504.

[3] A. Zielezinski, S. Vinga, J. Almeida, W. M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biology 18 (1) (2017) 186.

[4] G. Bernard, C. X. Chan, Y. B. Chan, X. Y. Chua, Y. Cong, J. M. Hogan, S. R. Maetschke, M. A. Ragan, Alignment-free inference of hierarchical and reticulate phylogenomic relationships, Briefings in Bioinformatics.

[5] Z. Xu, B. Hao, Cvtree update: a newly designed phylogenetic study platform using composition vectors and whole genomes, Nucleic Acids Research 37 (Web Server issue) (2009) W174–W178.

[6] M. Deng, C. Yu, Q. Liang, R. L. He, S. S. T. Yau, Correction: A novel method of characterizing genetic sequences: Genome space with biological distance and applications, Plos One 6 (3) (2011) e17293.

[7] G. E. Sims, S. R. Jun, G. A. Wu, S. H. Kim, Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions., Proceedings of the National Academy of Sciences of the United States of America 106 (8) (2009) 2677–2682.

[8] C. E. Shannon, A mathematical theory of communication, Bell Labs Technical Journal 27 (3) (1948) 379–423.

[9] J. Qi, B. Wang, B. I. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach., Journal of Molecular Evolution 58 (1) (2004) 1.

[10] A. J. Gentles, S. Karlin, Genome-scale compositional comparisons in eukaryotes, Genome Research 11 (4) (2001) 540–546.

[11] W. E. Johnson, E. Eizirik, J. Pecon-Slattery, W. J. Murphy, A. Antunes, E. Teeling, S. J. O'Brien, The late miocene radiation of modern felidae: a genetic assessment., Science 311 (5757) (2006) 73–77.

[12] U. Arnason, A. Gullberg, M. Janke, Akullberg, Mitogenomic analyses of caniform relationships, Molecular Phylogenetics & Evolution 45 (3) (2007) 863–874.

[13] Z. B. M. X. Y. B. Z. Z. . Z. F. Peng, R., The complete mitochondrial genome and phylogenetic analysis of the giant panda (ailuropoda melanoleuca), Gene 397 (1) (2007) 76–83.

[14] L. G, D. B, E. E, M. W, Phylogenomic evidence for ancient hybridization in the genomes of living cats (felidae)., Genome Research 26 (1) (2015) 1.

8. Figure Legends

Figure 1. The kmer degeneration (KD) approach. The analysis of $K = 7$ kmer for a 15 bp sequence of "ATGTCTAGTCTAGTC" was introduced as a case. **A**. The standard kmer analysis approach. The two red boxes show that the two kmers with different starting points in the sequence belong to the same kmer type. **B**. The degeneration of these kmers. Three positions with the combination of 1-3-7 positions were used to degenerate every kmer. For each kmer, only the 3 positions were maintained, with the other 4 position degenerated as "N". The results of the degenerated $K(N)$-mer (in this case the 3-7-mer) are shown on the right. With the KD, one more 3-7-mer became identical, shown in green. The two different original 7-mers were "TGTCTAG" and "TCTAGTG". With KD they degenerated as the same 3-7-mer of "TTG". Thus, the kmer type was deduced.

Figure 2. The distribution for 49951 balls into 65536. The horizontal axis is the number of balls in boxes. The vertical axis is the number of corresponding boxes.

Figure 3. The dispersed form versus the continuous form of $K(N)$-mer. **A**. The dispersed form in the case of $K = 7$ and $N = 3$. **B**. The continuous form in the case of $K = 7$ and $N = 4$.

Figure 4. Determination of the optimal $K(N)$-mer for a distribution. Panel **A** to **F**. Distribution of the Shannon information of the 1000 randomly sampled combinations for $N = 5, 6, 7, 8, 9$ and 10, respectively. For small N values, the distributions were scattered with multiple peaks. Since $N = 8$, the distribution became one peak.

Figure 5. Influence of the K value on the absolute value of Shannon information for the dispersed $K(N)$-mer form. **A**. Changes in the Shannon information value with K values for 10 different human genome segment sequences. Each color denotes one sequence. **B**. The positions chosen from the K-mer for the $K(N)$-mer. The K-mer lengths corresponded to those in panel **A**.

Figure 6. Performance of kmer degeneration employed for improving cvTree constructing phylogenetic tree of complicated mammal genomes **A**. phylogenetic tree of $K = 50$; **B**. phylogenetic tree of $K = 60$; **C**. phylogenetic tree by cvTree; **D**. *Feliformia* suborder referred from (Warren et al. 2012Gang Li et al. 2017).

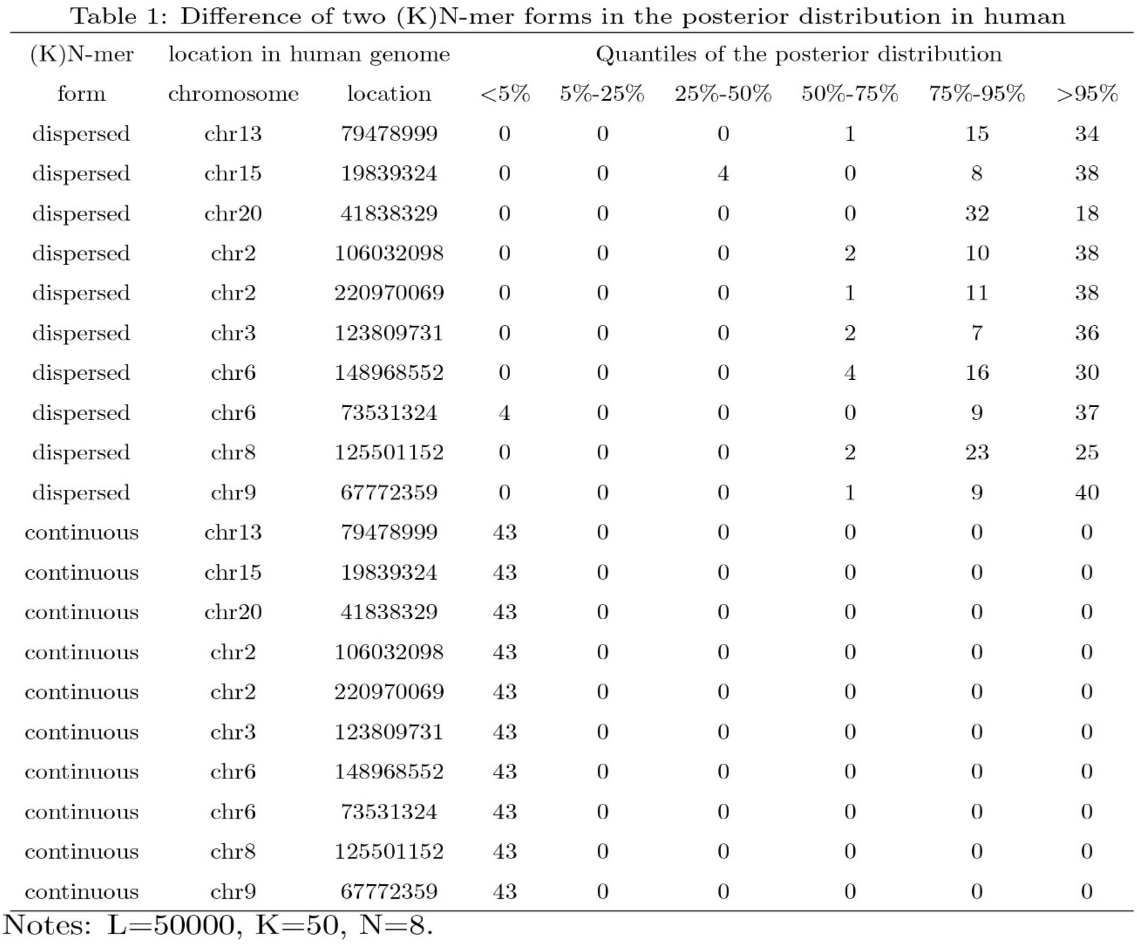


Table 1: Difference of two (K)N-mer forms in the posterior distribution in human

(K)N-mer	location in human genome		Quantiles of the posterior distribution					
form	chromosome	location	<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
dispersed	chr13	79478999	0	0	0	1	15	34
dispersed	chr15	19839324	0	0	4	0	8	38
dispersed	chr20	41838329	0	0	0	0	32	18
dispersed	chr2	106032098	0	0	0	2	10	38
dispersed	chr2	220970069	0	0	0	1	11	38
dispersed	chr3	123809731	0	0	0	2	7	36
dispersed	chr6	148968552	0	0	0	4	16	30
dispersed	chr6	73531324	4	0	0	0	9	37
dispersed	chr8	125501152	0	0	0	2	23	25
dispersed	chr9	67772359	0	0	0	1	9	40
continuous	chr13	79478999	43	0	0	0	0	0
continuous	chr15	19839324	43	0	0	0	0	0
continuous	chr20	41838329	43	0	0	0	0	0
continuous	chr2	106032098	43	0	0	0	0	0
continuous	chr2	220970069	43	0	0	0	0	0
continuous	chr3	123809731	43	0	0	0	0	0
continuous	chr6	148968552	43	0	0	0	0	0
continuous	chr6	73531324	43	0	0	0	0	0
continuous	chr8	125501152	43	0	0	0	0	0
continuous	chr9	67772359	43	0	0	0	0	0

Notes: L=50000, K=50, N=8.

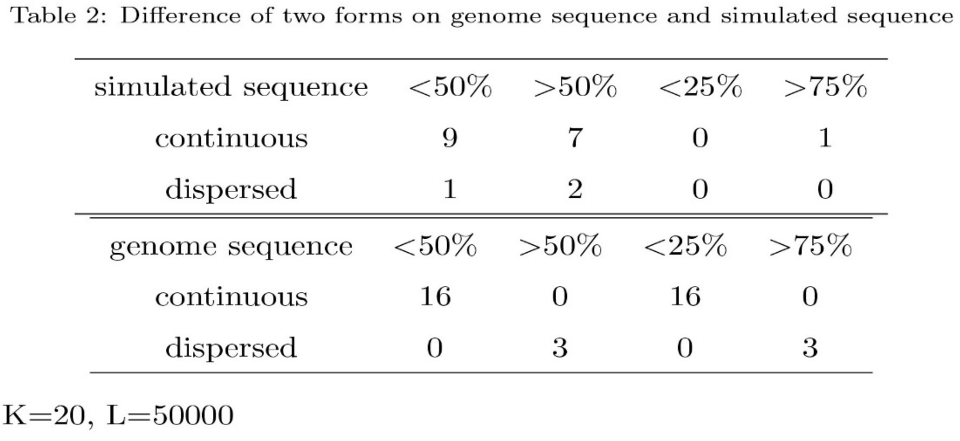


Table 2: Difference of two forms on genome sequence and simulated sequence

simulated sequence	<50%	>50%	<25%	>75%
continuous	9	7	0	1
dispersed	1	2	0	0
genome sequence	<50%	>50%	<25%	>75%
continuous	16	0	16	0
dispersed	0	3	0	3

N=5, K=20, L=50000

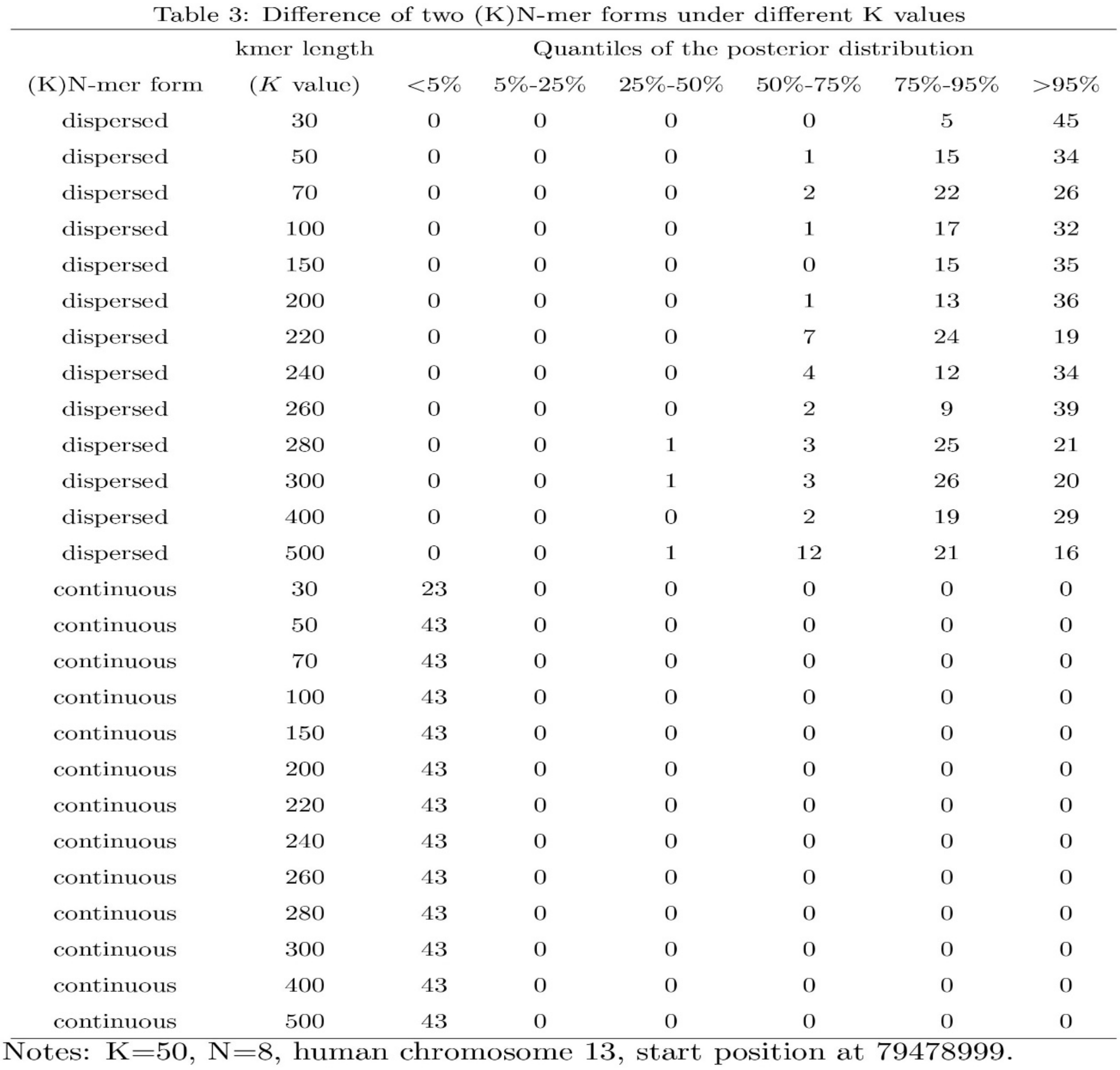


Table 3: Difference of two (K)N-mer forms under different K values

	kmer length	Quantiles of the posterior distribution					
(K)N-mer form	(K value)	<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
dispersed	30	0	0	0	0	5	45
dispersed	50	0	0	0	1	15	34
dispersed	70	0	0	0	2	22	26
dispersed	100	0	0	0	1	17	32
dispersed	150	0	0	0	0	15	35
dispersed	200	0	0	0	1	13	36
dispersed	220	0	0	0	7	24	19
dispersed	240	0	0	0	4	12	34
dispersed	260	0	0	0	2	9	39
dispersed	280	0	0	1	3	25	21
dispersed	300	0	0	1	3	26	20
dispersed	400	0	0	0	2	19	29
dispersed	500	0	0	1	12	21	16
continuous	30	23	0	0	0	0	0
continuous	50	43	0	0	0	0	0
continuous	70	43	0	0	0	0	0
continuous	100	43	0	0	0	0	0
continuous	150	43	0	0	0	0	0
continuous	200	43	0	0	0	0	0
continuous	220	43	0	0	0	0	0
continuous	240	43	0	0	0	0	0
continuous	260	43	0	0	0	0	0
continuous	280	43	0	0	0	0	0
continuous	300	43	0	0	0	0	0
continuous	400	43	0	0	0	0	0
continuous	500	43	0	0	0	0	0

Notes: K=50, N=8, human chromosome 13, start position at 79478999.

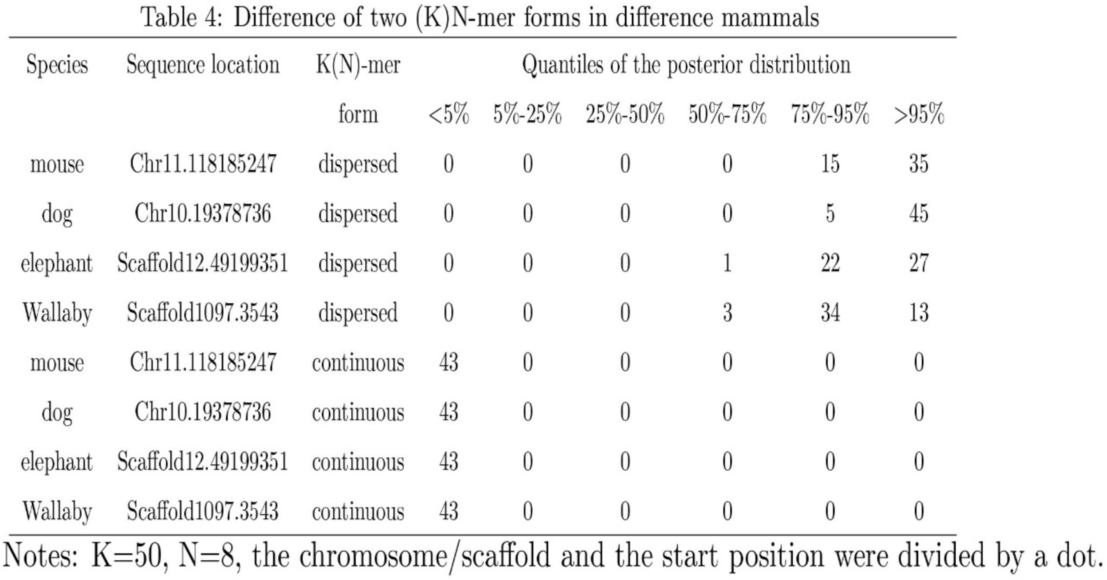


Table 4: Difference of two (K)N-mer forms in difference mammals

Species	Sequence location	K(N)-mer	Quantiles of the posterior distribution					
		form	<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
mouse	Chr11.118185247	dispersed	0	0	0	0	15	35
dog	Chr10.19378736	dispersed	0	0	0	0	5	45
elephant	Scaffold12.49199351	dispersed	0	0	0	1	22	27
Wallaby	Scaffold1097.3543	dispersed	0	0	0	3	34	13
mouse	Chr11.118185247	continuous	43	0	0	0	0	0
dog	Chr10.19378736	continuous	43	0	0	0	0	0
elephant	Scaffold12.49199351	continuous	43	0	0	0	0	0
Wallaby	Scaffold1097.3543	continuous	43	0	0	0	0	0

Notes: K=50, N=8, the chromosome/scaffold and the start position were divided by a dot.