Junwen Wang, Ph.D.

Research Interests

Overview

I am interested in developing novel algorithms and models to solve biological sequence related problems. The paradox we are facing today is the extreme complexity of biology and the extraordinary simplicity of its building units. DNA is made of only 4 bases and protein is made of 20 amino acids. However, there are thousands of genes and proteins in human cell that are built upon them, not to mention much larger number of interactions. To solve this puzzle we need to understand how these units are organized and how they interact to achieve their functions. I believe novel computer algorithms and models can better describe this dynamic process. Based on my previous research and knowledge, I intend to approach the problems in following four ways.

(Reference number corresponds to publication number in C.V.)

Exploiting contextual information to analyze biological sequences

Though the building units of DNA and protein are simple, the combinations of these units are more informative. For example, when we explore the helix propensity patterns using the neighboring amino acids, we found that the propensity for amino acid pairs is more significant and it was not always predictable by assuming proportional contributions from the individual propensity of the amino acids (**ref 3**). Under this observation, we reasoned that neighboring amino acids should interact more tightly with each other. We thus developed a protein sequence pairwise alignment algorithm that incorporated neighbor-dependent propensity. The algorithm is more effective in aligning distantly related or so called twilight zone sequences (**ref 1**).

The Markov model is a context dependent approach to model biological sequences. It calculates the probability of present unit based on the observation of its preceding units. The order of the model is defined as how many preceding units we look at. Generally, the higher the order (the more context we look), the better the model. However, higher order needs substantially larger training set and exponential computer time, which makes the implementation infeasible. Inspired by neighbor-dependent amino acid approach, we reasoned that neighboring nucleotides, like amino acid, should also influence each other more and should be considered together. Instead of using single nucleotide as model unit, we used di and tri nucleotide. In this way, 6^{th} order single nucleotide model can be replaced by 3^{rd} order di-nucleotide model or 2^{nd} order tri-nucleotide model and the computer time is significantly less.

We tested the performance of different Markov models to classify 10kb length 1,460 CpG-poor promoter sequences against 5,000 background sequences by 10-fold cross validation. We used 5^{th} Markov Model (MM) (Salzberg, 1998) as the baseline for which the average Correlation Coefficient (cc, a comprehensive measurement for classification accuracy, the high the better) was 0.1790. The corresponding figure for the 6^{th} order MM was 0.1983, a marginal improvement over 5^{th} order, which indicated that simply increase order will not do much better. We then attempted using di-nucleotide as model unit to capture joint dependence of neighboring nucleotides on the preceding ones. The corresponding figure for the 3^{rd} order di-nucleotide MM is 0.2713. We further use tri-nucleotide as model unit and the CC increase to 0.3244 on the same datasets. In conclusion, the by considering larger units for Markov models, we can improve both classification accuracy as well as the computer time (**ref 5**).

The context idea extends further to substitution matrices: we all knew that 20*20 amino acid substitution matrix is more discriminative than 4*4 nucleotide matrix for sequence alignments. We have good reason to assume that a 62*62 codon-based matrix will be better than amino acid

based matrix. The first generation of amino acid matrix, were derived from related protein sequences, when protein sequences were easier to obtain, and the large-scale genomic sequences were not available. However, genome sequences are abundant today, which enable us to construct a codon-based substitution matrix. I believe it is time to use more discriminative codon based matrix for sequence alignment. We did some preliminary work on this idea, constructed a codon matrix based on small dataset. We found it superior to BLOSUM62 matrix (**ref 8**).

Future direction*:* Based upon above experiments, the further question should be asked? What sequence feature other than the neighboring interaction should be considered and added to the model? We knew that both protein and DNA has periodicity characteristics. For example, protein helix structure is made of 3.6 amino acids repetition, it is reasonable to assume that the amino acid at the first helix position will have stronger interaction with the amino acid at the 4^{th} position, which is right above/below it in the helical structure. For DNA sequences, we knew that CpG di-nucleotide have functional relevance. Furthermore, analyzing a large number of bacterial promoters, a regular positioning of TA and TG stacks was detected with the best fit period of 5.6bp (Ozoline, 1999). The 5.6 bp period can be interpreted as a half of the DNA structure period of 11.2 bp, suggesting a sequence-dependent helical writhe of the promoter DNA. I will integrate these biologically meaningful remote interactions into algorithms or models in my future research. To make codon-based substitution matrix notable to public, substantially more work is needed to develop related software such as pairwise alignment, multiple alignment and phylogenetic tree building program based on codon matrix.

Discovering functional sub-regions by negative model

In contrast traditional method to search regions that have high score based on Markov model (trained from positive dataset), we instead searched for poor scoring regions, using a negative model (trained from "background" sequences). We found this approach is more effective to detect CpG island in human genome. We compiled a set of 1,319 CpG-poor promoters and randomly inserted an actual CpG island in each sequence. We then use a sliding window of 500 bps and score each window using the 5^{th} Markov model trained on the entire background sequences set (**ref 6**). When we used the valley score (the lowest score from background model) to discriminate 1,460 CpG-rich promoter and 5,000 background sequences we achieved a Correlation Coefficient of 0.9266 as compare to the values of 0.7330 for the 5^{th} order MM. In contrast, when we used peak score to discriminate the two datasets, we didn't get any improvements over 5^{th} order MM.

Future direction: this new way of finding interesting regions seems very promising and just like our previous enhancement, is applicable to many other functional signal detections such as first exon, MicroRNA(SiRNA) prediction, transcriptional module detection and gene recognization.

Considering positional restraint for functional site identification

There are many situations that we need to pinpoint some functional sites from biological sequences, such as Transcriptional Start Site (TSS), splice sites, HIV virus integration sites, etc. The cell can identify these sites efficiently, and we believe that there are signals around these sites and the process can be modeled and predicted.

The way we approached this problem is to use positional specific k-mer frequencies in combination with propensity analysis. We use a MM style algorithm that for each k-mer (k from 3 to 6), compute its frequency at each position. The frequency is converted to log-odds propensity score and put into the model. To score a sequence, we calculate a 200 bp window score by shifting one position at a time along the sequence. The window score is the sum of all the log odds propensity scores of k-mer at each position relates to the target site. We took the highest score as our prediction (**ref 7**).

We used 1,460 CpG-poor promoter sequences (1200 bp in length, upstream 1000 and downstream 200). We evaluated our method (PSPA) against 4 state-of-the-art methods (Bajic, 2004): PromoterInspector (Scherf, 2000), CorePromoter (Zhang, 1997), FirstEF (Davuluri, 2001), Dragon (Bajic, 2003). PSPA accuracy was based on a 10-fold validation experiment where the 1460 promoters were partitioned into 10 equal parts and we tested each part based on training on other 9 parts. Due to different requirement by the tools, we did two separate evaluations.

In first, the input was a 1200 bp region and the prediction correct if the predicted TSS is within ±50bp of the true TSS. Also due to licensing issues, we could only test 175 (randomly selected) promoters for PromoterInspector and Dragon. Table 1 summarizes the results. PromoterInspector did not make any prediction. Dragon made prediction only 26% of the times resulting in overall accuracy of 17%, but among the ones which were predicted, it correctly predicted 30 out of 46 (65%). When we only consider the top 26% of the PSPA predictions the accuracy within the predicted ones is 79%.

Program	#Sequences tested (1.2kb)	%Predicted	%Correct Prediction
PromoterInspector	175	0%	0%
Dragon	175	26%	17%
PSPA	146 x 10	100%	**47 ± 3%**

Table 1. Accuracy of TSS localization within 1.2 kb region

In the second way of evaluation, since FirstEF requires a larger context to make predictions, we used 10 Kb region flanking the TSS as the input. We call a prediction correct if the predicted TSS is within ±100 bp of the true TSS. Table 2 summarizes the results. FirstEF made a prediction only 23% of the times resulting in overall accuracy of 4%, but among the ones which were predicted, it correctly predicted 18%. When we only consider the top 23% of the PSPA predictions the accuracy within the predicted ones is 60%.

Program	#Sequences tested (10 kb)	%Predicted	%Correct Prediction
CorePromoter	1460	100%	2%
FirstEF	1460	23%	4%
PSPA	146 x 10	100%	**16 ± 2%**

Table 2. Accuracy of TSS localization within 10 kb region. Random expectation of accuracy is about 2%.

Future direction: PSPA is a powerful method for specific site identification. I will extend this method to splicing site, HIV virus insertion site, RNA A to I site localization. I will also build a PSPA-based transcriptional factor-binding sites library, to replace Transfac©. The PSPA-based model will have higher discriminating power than PWM since we take into account of history effects of each position.

Comparative genomics gives us more insights

Comparative genomics not only provide us several folds of data we can use for model development, but also gives us well conserved evolutionary signature that allow us to identify functional modules. We ask if the core-factor binding sites are more conserved between Human and Mouse than the surrounding region. We did the experiment with the CpG-poor datasets where the TSS was within a axtNet Human-Mouse alignment from UCSC database, using 4 core factors as shown in table 3. For each factor, the positive dataset consists of percent identity of the cis-elements (using the criterion for the sub-region and score-threshold corresponding to 5% false positive). The negative dataset consists of percent identity of randomly picked “sites” of same length from the conserved regions in the “surrounding” region. We use two definitions of “surrounding”: ±100bp and -1kb to +200bp flanking the TSS. Given the positive and negative sets of percent identities, we calculated the p-value using Wilcoxon’s paired test and the average sequence identity from both sets. As shown in Table 3, when we use ±100bp, TATA-box are significantly more conserved, but other core-factors do not show more conservation than the

surrounding region. However, when we used 1.2kb as negative set, all the core-factors are significantly more conserved. Clearly, the TSS immediate surrounding regions are more conserved than the distal regions.

Core element	Positive	Negative (±100bp)		Negative (±1.2kbp)	
	Avg %id	Avg %id	p-value	Avg %id	p-value
TATA box	75.54%	71.21%	5.46E-04	63.44%	4.24E-08
GC box	70.97%	71.36%	0.74	65.72%	3.88E-05
INR	71.76%	69.54%	0.15	65.00%	7.11E-03
CAAT	76.73%	72.75%	0.16	65.74%	5.42E-04

Table 3.Preferential Human-Mouse Conservation of core-factors

Future direction: the above result provides a very specific rational to use conservation as a criterion to reduce false-positive predictions, we will integrate the evolutionary conservation into our models and algorithms.

TB structural genomics

I am also collaborating with Dr. Bernhard Rupp at LLNL on TB structural genomics crystallization data exploration and pattern extraction. The initial stage data analysis has led to a publication in peer-reviewed journal (**ref 2**). A much large database will be available on 2005 and we are expecting more positive results.

Summary

The above mentioned methods are general approaches that can be applied to many area, such as gene identification, TSS identification, splice site identification, host-pathogen site prediction, RNA A to I editing site and protein family prediction. My short-term goal will be to further extend the application of the above algorithms to these areas. My long-term goal will focus on developing novel algorithms and models for biological sequence analysis.

Reference

Salzberg, S., Delcher, A., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Research* **26(2)**: 544-548.

Ozoline, O.N., Deev, A.A. and Trifonov, E.N. (1999) DNA bendability--a novel feature in E. coli promoter recognition. *J. Biomol. Struct. Dyn.* **16(4)**: 825-31.

Scherf, M., Klingenhoff, A., and Werner, T. (2000) Highly specific localisation of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. *J. Mol. Biol.* **297:** 599-606.

Zhang, M.Q. (1997) Identification of Human Gene Core Promoters in Silico. *Genome Res.*, **8**: 319-326.

Davuluri, R.V., Grosse, I. and Zhang M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genetics* **29**: 412-417.

Bajic,V.B. and Seah,S.H. (2003) Dragon Gene Start Finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.*, **13**: 1923–1929.

Bajic, V.B., Tan, S.L., Suzuki, Y. and Sugano, S. (2004) Promoter prediction analysis on the whole human genome. *Nature Biotechnology* **22(11)**:1467-1473.

Junwen Wang, Ph.D.

Teaching Interests

Bioinformatics itself is an interdisciplinary program that needs knowledge in biology, statistics, genetics, programming and database. It is my pleasure to teach courses that will allow students of all backgrounds to understand the different aspects of problems in the field. A major trend in biomedical science education now is that the proportion of quantitative science course, especially bioinformatics, needs to be substantially increased in student's curriculum. I started as a biologist, and later shifted to become a hardcore bioinformatics algorithm developer. I know what it takes for biology major to be a qualified bioinformatician. I would like to share the knowledge I gained through my various experiences and integrate them into my teaching.

While I am open to other subjects, I am mainly interested in teaching courses in bioinformatics/ computational biology. I would especially interest in an introduction course to bioinformatics, introducing common topics and databases, basic algorithms and models in bioinformatics. An introduction course to basic programming and data structure (such as introduction to java) that targets biomedical major would also be an option. Other higher level courses, such as topics in Markov models in gene identification and regulation, protein structure and functions, protein-protein interaction and networks are also possible.

I have strong desire to share my knowledge to students. I enjoy reading scientific publications and doing cutting-edge researches, I will convey the up-to-date information to my students through teaching. I also want feedback from students, I will find opportunity in my lab for students that have great idea and want to try.