Application of Analytic Number Theory in Bioinformatics

Jinrui Wang

Department of Mathematics, Shaanxi Xueqian Normal University, Xi'an, 710100

Keywords: Analytic Number Theory, Bioinformatics, Application Study

Abstract. Many topics in the field of bioinformatics can be abstracted into character sequence processing problems, such as gene recognition, protein secondary structure prediction, and so on. The sequence of characters can provide information from two aspects: composition and arrangement. The composition of the information can be used to reflect the conventional frequency. The key to the problem is how to reflect the arrangement of the character sequence. On the basis of summarizing the existing algorithms, this paper attempts to view the problem of character sequence analysis from the perspective of number theory, and puts forward the analytic number theory model of character sequence. In this model, the character sequence is regarded as a representation of the number, so that the problem of character sequence analysis is transformed into a number theory problem and solved by mathematical analysis method.

Introduction

The data is processed and processed to become useful information for people; information is refined and refined into a simple and understandable principle to guide people in practice before becoming theoretical knowledge. Research on how to achieve this process of discipline, called data mining or knowledge discovery. The growth of human-related knowledge is very slow compared to the biological data that is growing exponentially. On the one hand is the massive data, on the other hand is our medicine, medicine, agriculture and environmental protection and other aspects of the desire for new knowledge, with a view to improving their living environment and improve the quality of life, which constitutes a huge contradiction. This contradiction has greatly promoted the study of knowledge discovery in the field of biomedicine. We are witnessing a great change in the field of biomedical research, that is, from the traditional study of individual genes and individual proteins to systematic research on genomics, proteomics, and transcriptomics of whole organisms; Research methods from the traditional observation and experiment-oriented, transition to mathematics, information science and computer science and other disciplines theory and methods combined. This change makes a large number of mathematical science workers naturally transferred to the field of life science research, making bioinformatics this new discipline came into being. Since the birth of the late 80s, bioinformatics has taken the genome of informatics as the core, and has played an important role in the analysis of gene data, searching for new genes, analyzing and predicting protein structure function, molecular evolution and drug design and so on.

Biological Sequencing Engineering and Biological Sequence Database

With the implementation of the Human Genome Project, parallel research on the biological genome of microbes, plants, and animals can be used for the preparation of methodological and organizational work for human genome research. These studies are called "model bio-genome project" The Patterns of the Bio Genome Project The first identified patterns of organisms are: E. coli, yeast, Arabidopsis thaliana, nematodes, fruit flies and mice, etc., and then gradually added to some other species of model organisms such as puffer fish, zebrafish, etc. The In addition, some genettes of important production value and livestock and poultry, such as rice, wheat, domestic pig, chicken, silkworm and so on, are also added to the sequencing plan. The microbial genome is relatively small and closely related to the cause of health care, so the sequencing of the microbial genome is carried out in large quantities. The microbial genome to be sequenced mainly includes prokaryotes such as bacteria and archaebacteria, and lower eukaryotes such as fungi. This part of

the sequencing work, known as the Microbial Genome Project (MGP), was initiated by the US Department of Energy in 1986. The first fully sequenced creature in human history is a microbe - Haemophilus influenzae.

According to the statistics of the database GOLD (Genomes Online Database), as of September 25, 2003, more than 160 free organism whole genomes (including 4 chromosomes) have been completed, including 17 archaea , 127 bacteria and 20 eukaryotes. In addition, at least 410 important human pathogens, biologically significant and potentially economically valuable microbial genomes will be sequenced. At present, the sequencing of a large number of DNA sequence data, mainly stored in GenBank / EMBL / DDBJ and other international public nucleic acid sequence database, and published on the Internet, so that the world's relevant researchers to share.

Analytic Number Theory Model

The learning process of the dual descriptor is an unconstrained optimization process, which has the following two characteristics: (1) Strong uniform convergence. Since the mode deviation function is a quadratic function, it has a unique global minimum. Regardless of the initial parameters of the dual descriptor, the mode deviation function eventually converges to this unique global minimum. Because there is no local minimal interference, so the dual descriptor learning process is consistent convergence. (2) Convergence speed. As mentioned above, the minimum value of the mode deviation function is unique, it depends only on the characteristics of the sequence itself and the selection of the basis function, but not with the initial value of the dual description sub-parameter. However, the excellent dual descriptor corresponding to the minimum is not unique and it depends on the initial value of the parameter, which is the result of the pattern description function being designed as a two product. It is the uniqueness of the unique dual descriptor that gives it the ability to balance with it. Because there is no unique dual descriptor, the dual descriptor in the learning process, there is no need to travel long distances to a group of unique parameters, and as long as the initial value in the vicinity of the benefits, then find a the mode deviation function achieves a very small and extremely good dual descriptor on the line. Therefore, the convergence rate is faster.

The result of the above dual description sub-learning is to obtain a modal deviation from the function minimum and an excellent dual descriptor. So what's the use of them? In fact, the process of dual description of sub-learning is the process of approximating the best description of the sample. The result of the study is the feature of the sequence extracted from the description of the sample. For example, the learning process is to describe the sequence of tailored, and access to the excellent dual description is made of a dress. This dress is done in accordance with the sequence to be described, its size, fat just reflects the characteristics of the original sequence. Sequence recognition is to try to identify the sequence to try this dress, if appropriate, indicating that the sequence to be identified and the original sequence between the similar characteristics, if not appropriate, then the two sequences far between. In general, the use of dual descriptors is a common feature of a class of (multiple) character sequences, and the resulting excellent dual descriptor is a piece of clothing that reflects the common characteristics of the sequence of characters. If you want to identify the sequence wearing appropriate, you can put it in this category.

One-time study. Usually given a set of weight factors, and then according to the mode deviation function is very small, find the position weight function of an approximation I (k), then the position weight function to carry the sequence of feature information. For the character sequence s' to be recognized, the function is weighted by the function I (k), that is, the I (k) is introduced into the statistical process as a human disturbance. Assuming that the composition is equal, the permutation of the dual variable is extracted according to the duality formula, normalized as the feature quantity of the sequence. The method is: from left to right to scan the sequence to be identified s ', located in s' position k encountered character ic, then the character ic corresponding to the count variable increase * I (k), until the end of the sequence, and then The permutation variable ixI is divided by the length of the sequence s' as the feature quantity of the original sequence. Since the permutation part of the dual variable indicates the frequency of weighting, it is normalized to represent the

weighted frequency and can reflect the information of the composition and arrangement, so it can be used as the identification variable of the original character sequence. With the identification of variables, you can combine some other discriminant methods, such as Fisher discrimination, based on positive and negative samples to train the two parameters, the identification of the sequence.

The result of alternating learning corresponds to the global minimum of the mode deviation function. The excellent dual descriptor generated by it carries the feature information of the original sequence. For the sequence to be identified, as long as the description of this excellent dual descriptor from beginning to end, calculate its mode deviation from the function value, if the resulting d value is small enough, less than a preset threshold, you can put it and the original The sequence is homogeneous. The specific approach is: from left to right in turn to check the sequence to be identified s ', located in s' at the location k encountered character ic, with the character ic corresponds to the composition of the weight factor ix multiplied by the location weight function * I here (K), the resulting product is subtracted by 1 and then summed to the value d of the mode offset function, until the end of the sequence, and then the final d value and the preset threshold For comparison, if less than the threshold, then the sequence to be identified and the original sequence classified as similar, otherwise, different classes.

Application Examples in Bioinformatics

The extraction of gene coding loci and gene recognition are two aspects of a problem. The feature of the gene coding region is found, and it is determined whether or not it is a coding region depending on whether or not the sequence to be recognized has the characteristic. DNA sequence protein coding region recognition is commonly referred to as narrow sense of gene recognition, and broad sense of gene recognition refers to the integrity of the gene structure identification. As a result of the large number of genome sequencing projects, the international public nucleic acid sequence database DNA sequence data more and more, only by experimental methods to determine the gene and its location, both in time and money are unbearable. Thus, computer-aided genetic recognition has become an important topic in the field of computational biology (bioinformatics). The basic problem of computer identification (gene finding, or gene recognition) is to correctly identify the extent of the gene and its position in the genome sequence after a given genome sequence. After more than 20 years of efforts, dozens of predictive protein coding genes have been proposed, of which there are ten important algorithms and the corresponding software provides free online services.

Computer-aided genetic recognition algorithms can be broadly divided into two categories: gene recognition based on sequence homology and gene recognition based on sequence statistical features. Gene sequence recognition based on sequence homology, using a sequence alignment tool, such as BLAST or FASTA, to search for a known sequence in a nucleic acid sequence without a redundant database and to determine the sequence to be identified based on the homology size (degree of similarity) Of the gene position and function. For example, the identification tool ORPHEUS [136] is mainly based on sequence homology, taking into account the use of codons, the recognition tool CRITICA is mainly based on sequence comparison information, supplemented by six nucleotides.

At present, some of the problems in the identification of prokaryotic genes are: (i) the prokaryotic gene has no intron, but the gene interval is very small, the gene is prone to overlap (for example, overlap 4bp or 13bp), 5 ' It is difficult to predict correctly at the beginning. (Ii) the composition of short genes (eg, <60 ~ 80 amino acid residues) is not obvious and the statistical model is difficult to correctly identify. (Iv) For some genomes, especially those with high G + C (iii) the statistical model is overly dependent on the training set and the base is "atypical" (eg, horizontally transferred genes) Content of the genome, pseudo-positive rate is too high.For eukaryotic gene recognition, due to the presence of intron, its genetic structure than the prokaryotic complex, therefore, correctly identify the start codon, the codon and the complete gene structure are quite difficult. The existing eukaryotic gene recognition algorithm has a high recognition rate of 90% at the nucleotide level; however, the correct rate of recognition at the exon level is low , Less

than 50%. [148] .Therefore, the research status of the research is not satisfactory, there is still a lot of work to do. Gene recognition based on sequence homology and gene recognition based on sequence composition has their own advantages and disadvantages. For example, for a newly sequenced bacterial genome, only about 60 to 70% of the genes have homologous sequences in the current database, and about 30 to 40% are new genes that can not be found in known genetic databases. The recognition rate of the latter is relatively high, but there is an overly dependent on the training set, the pseudo-positive rate is high. If the two types of methods can be used in combination, complementary redundancy, gene recognition results may be better.For example, genetic recognition tools BDGF binding sequence Homology and statistical characteristics, and achieved good predictions.

Conclusion

The analytic number model of the character sequence, that is, the dual description sub-method, is based on the theory of character sequence analysis, the idea of the perturbation method into the statistical field, the position-weighted frequency to reflect the composition of the character sequence and arrangement of two aspects of information, Is a general way to deal with character sequence problems. It does not limit the application itself, as long as it is involved in the problem of character sequence analysis, you can try to apply the method to solve. However, in recent years, due to the development of sequencing work, bioinformatics has produced a large number of biological sequence data, just for the method provides a useful. On the other hand, the model was originally proposed to be needed.

Acknowledgements

Shaanxi Pre-school Teachers College Research Fund Project (2015YBKJ072);

References

[1] Yang Yanling, Wang Jihua, Liu Huilan. Z-curve analysis of avian influenza virus in different hosts [J]. Journal of Dezhou University, 2008 (06)

[2] Wang Lan, Chen Jing, Wang Rui, Lu Xiaoquan. Application of several pattern recognition methods in bioinformatics [J] .Computer & Applied Chemistry, 2007 (01)

[3] Yang Yanling. A New Method for Identifying Genes-Z Curve [J]. Journal of Chifeng University (Natural Science Edition). 2007 (03)

[4] Liu Chao, Ma Zhiqiang, Liu Shuai. Dual Sequence Alignment Algorithm in Bioinformatics [J]. Journal of Changchun Institute of Technology (Natural Science Edition), 2007 (03)

[5] Ma Jianghong, Zhang Wenxiu, Xu Zongben. Data mining and database knowledge discovery: statistical point of view [J]. Journal of Engineering Mathematics, 2002 (01)

[6] Zhang Chunting. Geometric analysis of DNA sequence [J]. China Science Foundation, 1999 (03)