Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 2 Number 1 (2009) pp. 13–19

http://www.ripublication.com/jcib.htm

Computational Motif Signature Discovery and Validation of Gene Prediction in *Aspergillus terreus* NIH 2624

K Palani Kannan*, TK Subazini, CP Rajadurai, S Naga Vignesh and G Ramesh Kumar

AU-KBC Research Centre, MIT Campus of Anna University Chennai, India
**E-mail: kpalanikannan@au-kbc.org*

Abstract

Motif signatures are amino acid sequences that can be a part of domain of the protein and it plays a major role in the conformational arrangements of protein structures. In this research, a new, comprehensive method for the identification of motif signatures in the genome of *Aspergillus terreus* has been introduced. With the availability of complete *Aspergillus terreus* genomic sequences comprises 8 chromosomes and 10406 coding sequences, it is now possible to use computational methods to identify motif signature sequences, and to use these signatures as the basis for diagnostic assays, biomarker studies and to detect pathway analysis and genotype of the *A.terreus* in both commercial application wise and clinical application wise. The success of such analysis critically depends on the methods used to identify motif signatures that properly differentiate between the target sequences. We have used hidden markov model tool (Glimmer HMM) to compute accurate coding regions of the *A.terreus* genome by retraining the HMM tool and fingerPRINT scan for the motif signature prediction. A genomic coding sequences predicted from the trained genomic sequences has been successfully tested across PRINTS motif signature database, and the results indicate that the signatures having functional role. Those motif signatures predicted sequences are validated as new coding sequences in *A.terreus* genome and none reported yet.

Keywords. Hidden Markov Model, Motif signature prediction, *Aspergillus terreus*

Introduction

Modern computational and genomic data analysis concerns have raised interest in the real-time detection and identification of new coding sequences and signatures involved in the gene expressions. Microbes and low level eukaryotic fungus have always represented one of the greatest threats to human health and buzz in industrial applications, and in recent times this buzz in microbial applications increased due to the possibility of engineered biological agents. For these reasons, the genome sequencing field has targeted and sequenced the complete genomes of various prokaryotic microbes and eukaryotes over the past years, with many more new coding sequences expected to appear in the near future. These new sequences now make it possible to develop diagnostic assays capable of identifying organisms in environmental and clinical samples as a biomarker. Those sequences play a major role in the expression study of the proteins through pathway analysis and the biomarker studies. Those sequences play a role to distinguish the target application in the specific organism from all other known organism. A sequence that accurately distinguishes a target protein expression is termed as signatures. In general, a signature sequence must be conserved among a set of target genomes and dissimilar to any sequence in the surrounding environment. To detect target signatures, computational methods has been introduced using algorithms such as Hidden markov model and fingerPRINT scan.

Aspergillus Terreus

A. terreus NIH 2624 is filamentous and ubiquitous fungi found in nature. It is commonly isolated from soil and plant debris. In the microscopic analysis, the colonies of *A. terreus* on potato dextrose agar at 25°C are beige to buff to cinnamon [5]. It's growth rate is moderate when compared with other *Aspergillus. spp* and it produces lovastatin in the chemical medium that containing glucose and glutamate or histidine [8]. The hyphae of the *A. terreus* looks like septate and hyaline structures and it's colonies produce finely granular conidial spores. Conidial heads are biseriate and columnar. Conidiophores are smooth-walled and hyaline, 70 to 300µm long, terminating in mostly globose vesicles. Conidia are small (2-2.5 µm), globose, and smooth [3]. The *A. terreus* genome was sequenced by the "Broad Institute of Massachusetts Institute of Technology, USA" from the genomic DNA given by Dr. David Denning, University of Manchester. The whole genome was classified into 26 supercontigs (The whole genome segregated into several parts based on the higher length to lower length cut) as 267 contigs (The supercontigs segregated into several parts based on the higher length to lower length cut) and sequenced using whole genome shotgun methodology. The nucleotide sequences are shattered into small fragments (~4 kb or ~40 kb) and each fragment is inserted into vector elements and cloned. The two ends of the fragment are sequenced and paired reads was created. The assembly process uses the paired reads to identify contiguous stretches of sequence (contigs). Contigs are ordered and linked together into larger supercontigs using paired reads lying in different contigs. These predictions have 15 structural RNAs and 10406 coding sequences. The whole genome was sequenced on 02 Oct 2006 [6].

Computational Pattern Discovery

In Computational Biology, mathematics plays a key role and act as a back bone for the development. The gene and the motif signatures are kind of patterns and pattern prediction through mathematical concepts are evolving. The prediction of patterns using mathematical algorithms are efficient for the large number of data when compare with the other methods like homology based pattern prediction. The major concepts used in the mathematical gene prediction are normally probability based gene prediction methods like Hidden markov model [4] and fingerPRINT scan [9]. The computational researches those involved in the discovery of novel patterns from the large size genome adopt the mathematical gene prediction methodologies because of its probabilistic prediction of computationally intense new patterns. In connection with above criteria, *A. terreus* genome is a large data and the coding pattern regions and motif signatures available are very less. The mathematical pattern prediction strategy only can do the gene prediction and motif signature prediction for large data like *A.terreus* genome [10].

Computational Strategies and Algorithms

Glimmer HMM is generalized Hidden Markov Model implementations for *ab initio* eukaryotic gene prediction developed along with the specialized application of TigrScan. The C/C++ source code for both is available as open source and is highly reusable due to their modular and extensible architectures. Among the currently available genefinding tools, the programs are re-trainable by the end user. The special application on this tool are re-configurable and include several types of probabilistic submodels which can be independently combined, such as Maximal Dependence Decomposition trees and interpolated Markov models. These sub models are used in the case of multiple coding region prediction in the computational gene prediction of new raw sequence of the unknown genome. Both programs have been used at TIGR for the annotation of genes [7]. GlimmerHMM was analysed by the development team by performance test on the *A.fumigatus* set for three of the measures. The greater difference in accuracy between other tools on the *A.fumigatus* set demonstrates the value of being able to retrain the gene finders for specific organisms with respect to the quality of the training data set. Training is the process of developing the model for the mathematical machine learning programs to predict the genes in the given nucleotide sequence. The efficient training was performed using the large number of training set data. The efficiency of model development in the training process was directly proportional to the size of the training data.

FingerPRINT scan [9] is an algorithm used to identify the signature fragments of query sequences that match a database of motif signatures. A motif signatures are the sequential arrangements of amino acid residues excised using multiple sequence alignment using motif profile. The algorithm is the combined approach of Gribskov profile, BLOSUM and Dayhoff's matrices [9]. By repeated scoring of query sequence with the PRINTS database, it will give the aligned signatures. This algorithm we can use for the validation of gene predicted as well as the motif signature identification.

Material and Methods

The whole *A.terreus* genome supercontig fasta files and coding regions GFF (General Feature Format) files also obtained from the BROAD institute of MIT and Glimmer HMM was trained. From the training data, the Hiddem Markov Model program creates the trained model and active run process was performed based on the trained model. The active run process predicts coding regions and the predicted regions were analysed across the nt database with the help of Batch BLAST program [11]. The hit sequences and no hit sequences were segregated using the perl programs developed. From the no hit sequences, start and stop codons and fasta files for each coding sequences were generated manually using the perl programs. The fasta files were converted in to amino acid sequences using the codon table available for the *A.terreus*. The most of the sequences were short sequences, and it should be a motif signatures. Those amino acid sequences were analysed across the PRINTS database [1] using the fingerPRINT scan algorithm. The whole work is explained as flow in (figure. 1). The results from the fingerPRINT scan were analysed and discussed in following chapters

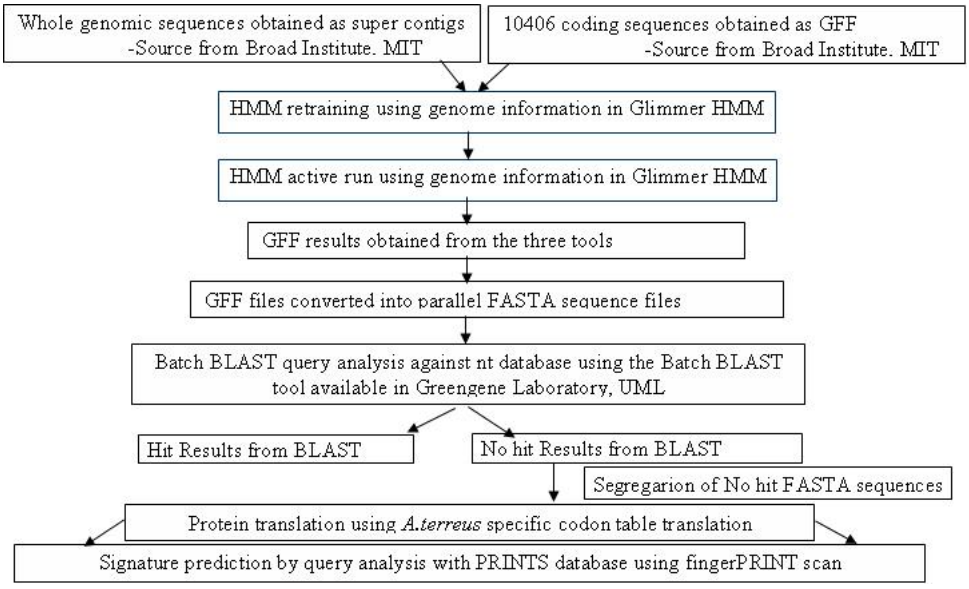

Figure 1: Work Flow.

Results and Discussions

Sequences already reported in the public databases were predicted by Broad Institute by the combined tools annotation process using FGENESH, GENEID and GENEWISE are available in NCBI genome database as – 10406 coding regions based on their strategy computational of gene prediction (Cited through Broad Institute Wesbsite [2]). The most of the computationally predicted sequences were named as hypothetical, conserved hypothetical and related sequences. So, the results published

cannot be exact coding regions. The results of this computational research work provide the sequences of 10406 sequences with added sequences 11486 as 21792 sequences (figure. 2). The increase in our prediction shows that some new novel regions were occurred in the predicted sequences. With these results, it is not confirmed that above predicted sequences are novel sequences.

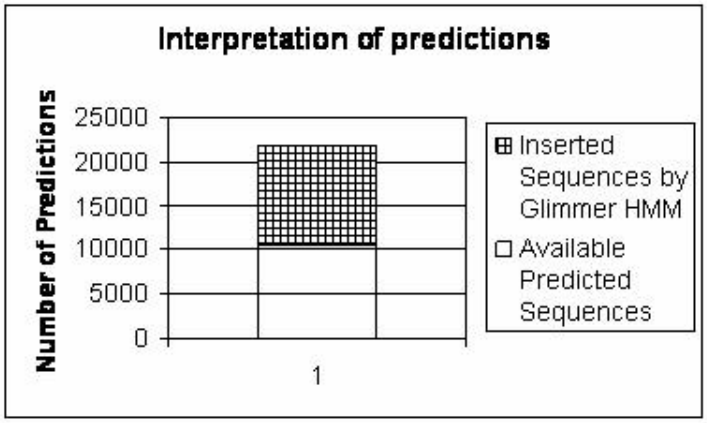

Figure 2: Comparative analysis of our prediction with respect to available sequences.

Thus, 21792 sequences were BLAST [11] with the nt (nucleotide) database for hit score analysis. The no hit score indicates that the query sequence have no similar hit in the nt database and none reported yet. 1995 sequences were reported as the no hit sequences (figure. 3). The no hit sequences were translated into amino acid sequences and then analyzed with the PRINTS database using fingerprint. It clearly indicates that the no hit sequences have the signature values.

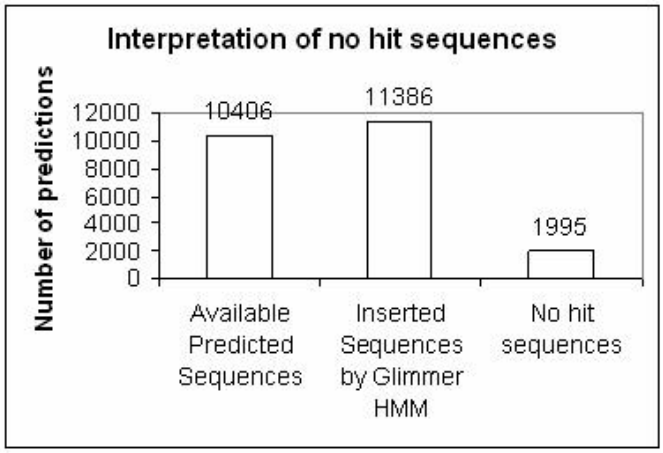

Figure 3: Comparative analysis of no hit prediction in BLAST analysis.

The fingerPRINT scan of 1995 no hit sequences shows that 866 sequences were signature hit in PRINTS database. Among the 866 motif signature predicted, B81Histone H2A signature, DNA topoisomerase II family signature I, Glucose-6-phosphate isomerase signature and Metalloprotease inhibitor signature as enzymes signatures as a featured prediction because of their role in the basic biochemical pathways like Glucose Metabolism and DNA replication.

Conclusion

In the case of *A.terreus* coding sequence predictions, the sequences were predicted by means of comparative analysis of various mathematical gene prediction results from various computational tools. If we study in depth to the gene predictions, every gene prediction tool will give the inserted sequences and missed sequences, so we can't confirm the occurrence of coding regions by computational efficiency by comparing the tools. The unique predictions as inserted sequences of every tool should be considered for the functional analysis. The inserted 11386 sequences of our prediction indicate that those sequences have a functional value as motif signatures and 866 sequences have motif signature hit. Those 866 nucleotide sequences were not reported by anyone and it was validated across nt database and considered those sequences as novel sequences. Among the 866 motif signature predicted, B81Histone H2A signature, DNA topoisomerase II family signature I, Glucose-6-phosphate isomerase signature and Metalloprotease inhibitor signature as enzymes signatures as a featured prediction because of their role in the basic biochemical pathways like Glucose Metabolism and DNA replication. Other motif predictions show various novel motif signatures. These new predictions may play a role in the metabolic pathway development and reconstruction for *A.terreus* as a biomarker. Those 866 sequences had been taken into the analysis of pathway development in lovastatin biosynthesis. Those 866 sequences may fill the pathway holes in the lovastatin biosynthesis pathway.

References

[1] Attwood T.K., Beck M.E., Bleasby A.J. and Parry-Smith D.J. 'PRINTS a database of protein motif fingerprints', (1994), Nucleic Acids Research, Vol. 22, No. 17, pp. 3590 -3596.

[2] Aspergillus Comparative resources, BROAD institute, MIT Harvard.

[3] Deanna S.A., Stephen S.E., Sanjay R.G., Annette F.W. and Michael R.G. 'In Vitro Amphotericin B Resistance in Clinical Isolates of *A. terreus*, with a Head-to-Head Comparison to Voriconazole', (1999), J. Clin. Micro. Bio., Vol. 37, pp. 2343–2345

[4] Ethem A. 'Introduction to Machine Learning', (2004), MIT Press, pp. 305-325.

[5] Hassan H., Peter N. and Philippe D. ‘Lovastatin Biosynthesis by *A. terreus* in a Chemically Defined Medium’, (2001), Applied and Environmental Microbiology, Vol. 67, pp. 2596–2602.

[6] Kim P.D., Tatiana T. and Donna M.R. ‘NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins’, (2005), Nucleic Acids Research, Vol. 33, pp. D501–D504.

[7] Majoros W.H., Pertea M., Salzberg S.L. ‘TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders’, (2004), Bioinformatics, Vol. 20, pp. 2878–2879.

[8] Vederas J.C., Moore R.N., Bigam G. and Chan K.J. ‘Biosynthesis of the Hypocholesterolemic agent Mevinolin by *A. terreus*’, (1985), J. Am. Chem. Soc., Vol. 107, pp. 3694–3701.

[9] Philip S., Darren R.F. and Teresa K.A. ‘FingerPRINTScan: intelligent searching of the PRINTS motif database’, (1999), Bioinformatics. Vol. 15, No.10, pp. 799-806.

[10] Richard D., Sean R.E., Anders K. and Graeme M. ‘Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids’, (1999), Cambridge University Press, pp. 1-77.

[11] Stephen F., Altschul P., Warren Gish, Webb Miller, Eugene W. Myers and David J. Lipman ‘Basic Local Alignment Search Tool’, (1990), J. Mol. Bio., Vol. 215, pp. 403-410.