SP 5240 P. 167
2011
SP-PP-5240

R.A.

estimated standard deviation was only 3%. Hence, this technology together with standardized bioinformatic analyses is suitable for a clinical diagnostic setting.

Poster U17

A systematic and agnostic learning method for identifying sequence motifs relevant to protein subcellular localization

Kuo-Bin Li *National Yang-Ming University*
Shou-Cheng Yen (National Yang-Ming University, Institute of Biomedical Informatics);

Short Abstract: This poster describes a novel method for identifying sequence motifs to predict protein subcellular localizations. Most existing methods rely either on prior knowledge about protein targeting signals or on sophisticated residue compositions that often don't provide clear insight. We proposed a systematic approach to identify signature motifs without using prior knowledge. We concentrated on the localizations that are traditionally more difficult to predict. For proteins within those localizations, we investigated all sequence motifs (length < 8) represented by a reduced amino acid alphabet set. Each motif was then subject to a statistical test to determine if it has a distinct occurrence frequency for proteins in a specific localization. The identified sequence motifs were further extended on both ends to increase their length, resulting in eight motifs for five localizations. Three of the motifs have never been applied to the prediction of localization, they are: (1) the [WFY][AVLI][AVLI]KNS[WFY] motif, a lysosomal specific motif found on cathepsin protease active site; (2) a RERXXER motif exclusive for peroxisomal proteins; and (3) an enriched CGHC motif present exclusively in ER proteins. The results facilitate implementations of more accurate prediction tools for lysosomal, peroxisomal and ER proteins, the three challenging localizations. With extension of proteins located in other subcellular compartments using a wider range of physicochemical properties, our discovery-oriented approach fulfills the gaps left by the current studies in this field.

Poster U18

A dotplot method for motif discovery: the importance of outlier alignments in the score-abundance relationship.

Kazuhito Shida *Tohoku University*

Short Abstract: Basically, De novo motif discovery algorithms seek the gapless-alignments with unproportionally high alignment score for their statistical abundance expected in given background model.
However, precise determination of maximally "unproportional" alignment requires calculation of precise P-value under background models for numerous alignments, that is rarely feasible.

An extensive numerical experiment reveals that a special definition of score and abundance are in a simpler relationship than definitions used in typical motif analysis.
In this special definition, the mean alignment score is roughly proportional to the alignment abundance, which means the most statistically unusual alignment (possibly the biologically correct motif) can be approximately detected by a basic linear classifier.

This discovery method is even more simplified, implemented in Perl5 and tested on human portion of benchmark proposed by Tompa et al.
After a preprocessing by RepeatMasker, every W-mer in the input is used as a "seed" for a heuristics method to obtain locally optimal alignments under a ZOOPS-like occurrence model.
Then, using above-mentioned special definitions, the abundance and alignment score are evaluated for these selected alignments to be organized in a 2D dot-plot.
Finally, the outlier dot toward upper left corner is manually picked as the end result.
Surprisingly, this very primitive method performs at least as accurately as the best algorithms reported (MEME3, Weeder, YMF, etc.).

Also being developed are an automated outlier detection, automated parameter (e.g. W) setting, and a stochastic version (re-weighted Gibbs-sampling) of this method. An open-source version and a manuscript are under preparation, too.

Poster U19 ✓

ZEBU GENOME SEQUENCING AND ANALYSIS USING SECOND GENERATION TECHNOLOGY

Guilherme Oliveira *FIOCRUZ*
Adhemar Zerlotini (FIOCRUZ) Adhemar Zerlotini (FIOCRUZ, CEBio); Flávio Araújo (FIOCRUZ, LPCM); Betânia Drumond (UFJF, Virology); Izinara Rosse (UFMG, Genetics); Sara Cuadros-Orellana (FIOCRUZ, CEBio); Beatriz Lopes (ABCZ, ABCZ); Elizângela Guedes (EMBRAPA, CNPGL); Wagner Arbex (EMBRAPA, CNPGL); Marco Antônio Machado (EMBRAPA, CNPGL); Maria Gabriela Peixoto (EMBRAPA, CNPGL); Rui Verneque (EMBRAPA, CNPGL); Marta Guimarães (EMBRAPA, CNPGL); Roney Coimbra (FIOCRUZ, CEBio); Maria Raquel Carvalho (UFMG, Genetics); Marcos Vinícius Silva (EMBRAPA, CNPGL); Guilherme Oliveira (FIOCRUZ, CEBio);

Short Abstract: The Brazilian cattle are composed mainly by zebu breeds and crossbreeds between taurine and zebu breeds. During the last decades in Brazil, traditional genetic evaluations have guaranteed considerable gains in dairy production and resistance to diseases and parasites. Although effective, these methods do not shed light at the biological processes underlying the observed results. The inclusion of genetic markers in the breeding programs would produce 30 to 50 percent more genetic gain than a traditional progeny test system at a similar or lower cost. Aiming to identify genetic polymorphisms in the zebu genome of dairy Gyr breed, we constructed mate-paired genomic libraries with 1-2 kb inserts. Fifty bp-long reads produced with SOLiD V3 platform were mapped into the reference genome of a female Bos taurus (NCBI Project ID: 10708) using BioScope. SAMTools was used to generate the consensus sequence and to identify SNPs. The first two SOLiD v.3 runs yielded 204 million 50 bp-long reads representing ~1.18X observed coverage of the reference genome. An initial comparative analysis was performed for six genes related to dairy production and we observed low identity in coding regions compared to their Bos taurus orthologs. Twenty-one new SNPs were identified on the chromosomes that harbor

SP 5240
p 167

these genes, especially on chromosome 6 (19 SNPs), mainly on non-coding regions, except for one within Fibulin-7 gene on chromosome 11. The identification of SNPs in dairy Gyr cattle may improve the efficiency of the next version of genotyping chips to be used for Dairy zebu genomic selection in Brazil.

Poster U20

Improved performance of sequence search algorithms in remote homology detection

Adwait Joshi *National Centre for Biological Sciences (TIFR)*
Ramanathan Sowdhamini (National Centre for Biological Sciences (TIFR))

Short Abstract: Remote homology detection from mere sequence information is applicable for the analysis of genome databases, but is highly challenging due to sequence dispersion within protein families. Iterative profile-based sequence search algorithms and methods that employ Hidden Markov Models are quite effective in detecting remote homologues, however they seldom achieve full coverage. In this study, we have compared two such methods, iterative profile-based searches - Position Specific Iterative BLAST (PSI-BLAST) and a motif-initiated constrained profile-based search : Pattern Hit Initiated BLAST (PHI-BLAST). We have integrated various strategies for achieving high coverage including multiple queries (for PSI-BLAST) and multiple motifs (for PHI-BLAST). We have tested the strategies over 12 protein structural superfamilies present in PASS2 database (Bhaduri and coworkers, BMC Bioinformatics, 5, 35), which directly corresponds to SCOP but includes members with <40% mutual sequence identity. Following these search strategies followed by validation using Hidden Markov Model library of PASS2 superfamily members, the coverage at superfamily level was analyzed. Sequence searches driven by multiple motifs per query through PHI-BLAST clearly outperform PSI-BLAST performance suggesting its utility in better coverage in remote homology detection. The findings reveal saturation of number of homologues obtained for multiple query multiple motifs approaches through PHI-BLAST as compared with multiple query approach of PSI-BLAST followed by best performing query in both the methods. Whereas this must be the best approach, owing to the computational expense, a best performing query with its multiple motifs employed through PHI-BLAST can mitigate trade-off between coverage and computational time.

Poster U21

New method of detecting pathogenic viruses using high-throughput sequencing data

Daisuke Komura *The University of Tokyo*
Shumpei Ishikawa (The University of Tokyo)

Short Abstract: Massively parallel sequencing technology enables us to efficiently discover new viruses such as clonal infection of polyomavirus in Merkel cell carcinoma . Conventional methods assume that such viruses belong to known virus family, so that they would share closely related sequences. However, this method will overlook novel pathogenic viruses which do not have such sequences.

We present here a method to discover novel pathogenic viruses from sequencing data by searching sequences homologous to human. The method is based on the fact that some pathogenic viruses have genes homologous to humans, and such genes play important roles in pathogenesis. For example, some retroviral and human oncogenes are homologous to each other, and viral IL-6 protein encoded in human herpesvirus 6 facilitates neoplastic cell replication in Kaposi sarcoma. Simply searching for human homologous sequences may lead to many false positives due to sequencing or mapping errors. In order to avoid this, we utilize information on pathogenic context, pathways, and protein domains. Our method will enable us to identify new pathogenic viruses and the genes related to pathogenesis simultaneously. We evaluate usefulness of our approach on actual data set from human disease tissues.

Poster U22

Perfect Hamming Code as Hash key for Fast Genome Mapping

Yoichi Takenaka *Osaka University*
Shigeto Seno (Osaka University, Bioinformatic Engineering); Hideo Matsuda (Osaka University, Bioinformatic Engineering);

Short Abstract: This poster is based on Proceedings Submission 61.

With the advent of next-generation sequencers, the growing demands to map short DNA sequences to a genome have promoted the development of fast algorithms and tools. The tools commonly used today are based on either a hash table or the suffix array/Burrow-Wheeler transform (BWT). These algorithms are the best suited to finding the genome position of exactly matching short reads. However, they have limited capacity to handle mismatches. To find n-mismatches, they requires O(2^n) times the computation time of exact matches. Therefore, acceleration techniques are required.

We propose a hash-based method for genome mapping that reduces the number of hash references for finding mismatches without increasing the size of the hash table. The method regards DNA subsequences as words on Galois extension field GF(4) and each word is encoded to a code word of a perfect Hamming code. The perfect Hamming code defines equivalence classes of DNA subsequences. Each equivalence class has a representative subsequence and all the 1-mismatch subsequences from the representative belong to the class.

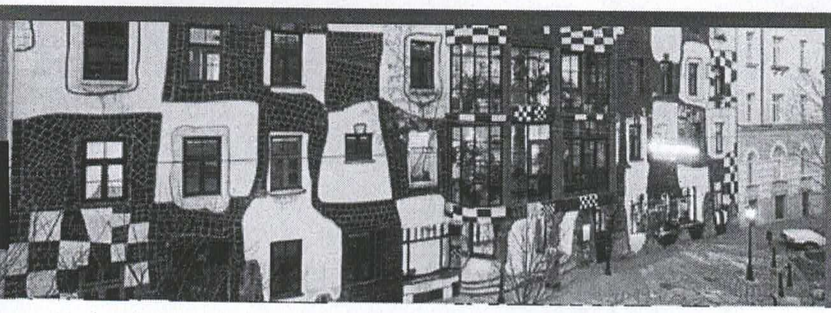
The code word is used as a hash key to store these subsequences in a hash table.
Specifically, it reduces by about 70% the number of hash keys necessary for searching the genome positions of all 2-mismatches of 21-base-long DNA subsequence. This method can also apply to BWT-based genome mapping.

Poster U23

Patome: a database for biological sequence annotation in patents

Byungwook Lee *Korea Research Institute of Bioscience and Biotechnology*

VIENNA
ISMB
ECCB
2011



An Official Conference of the International Society for Computational Biology

ISCB

July 15–16
SIGS AND TUTORIALS
July 17–19
CONFERENCE

search...

19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology

JOIN ISCB KEY DATES NEWS REGISTER

- HOME
- GENERAL INFO
- PROGRAM
- SPONSORS
- SUBMISSION
- COMMITTEES
- CONTACT
- PROMOTION
- PRESSPASS

ISCB Student COUNCIL SYMPOSIUM 2011

SIGs and Satellite MEETINGS

Thank You for Another Great ISMB/ECCB Conference!

See you next year at **ISMB 2012** in Long Beach, California (July 15-17), and at **ECCB 2012** in Basel, Switzerland (Sept 9-12).

Conference Press

Several journalists were in attendance at the conference. See what they wrote about here

Pictures

View ISMB/ECCB pictures on flicker: http://www.flickr.com/groups/ismbeccb2011

Tag your own conference pictures using the tage ismbeccb2011

Online ISMB/ECCB Conference Proceedings available at: http://bioinformatics.oxfordjournals.org/content/27/13

Awards

Several presenters were honored with awards at the conference. See which awards were given to whom here.

ISMB/ECCB 2011 was pleased to present the following keynote speakers.

- **Michael Ashburner, University of Cambridge, ISCB Senior Scientist Accomplishment Award**
- **Bonnie Berger, Massachusetts Institute of Technology**
- **Luis Serrano, Centre for Genomic Regulation**
- **Janet Thornton, European Bioinformatics Institute, ECCB 10th Anniversary Keynote**
- **Olga Troyanskaya, Princeton University, 2011 Overton Prize Award Winner**
- **Alfonso Valencia, Spanish National Cancer Research Centre (CNIO), ISCB Fellow Keynote**

Keynote details are available here